
Underspecification: from Semantics to Discourse — Invited Talk —

MARKUS EGG [†]

Underspecification has been introduced into semantics as a means to handle ambiguity. In the meantime, a host of underspecification formalisms is available, which represent the meaning of an ambiguous expression in terms of partial semantic information. Two properties of formalisms have emerged as crucial:

First, is a formalism *efficient*, i.e., can the readings of an ambiguous expression be derived and enumerated easily from its underspecified representation, even for very high numbers of readings? (In NLP applications, the number is much higher than one would expect due to spurious ambiguities, see Koller and Thater 2006). Second, is a formalism *expressive*, i.e., can it represent any subset of the readings of an ambiguous expression (König and Reyle, 1999, Ebert, 2005)?

Practical work on discourse annotation in Potsdam (Reitter and Stede, 2003) and in Groningen shows that underspecification is desirable for discourse processing as well, because not every discourse can be assigned a single fully specified structure by human analysts (or discourse parsers), which introduces ambiguity at the discourse level.

But for discourse processing, efficiency and expressivity of underspecification formalisms get even more important: The items to be analysed get drastically larger (the number of atomic segments in a discourse exceeds the number of scope-bearing entities in a sentence by far), which calls for much more efficient processing. And, discourse processing requires a high grade of

[†]This talk presents joint work with Michaela Regneri and Alexander Koller.

expressivity to allow the integration of preferences. These preferences can be extracted from large corpora annotated for discourse structure (for the corpora see Carlson et al. 2003 and Stede 2004).

The preferences describe the interaction of discourse relations (e.g., CONDITION, BACKGROUND, or SUMMARY) and discourse configuration (how smaller segments of discourse are arranged into larger ones), which together constitute discourse structure. Consider for instance the discourse (1):

- (1) I try to read a novel (C_1) if I feel bored (C_2) because the TV programmes disappoint me (C_3) but I can't concentrate on anything. (C_4)

For (1) five different discourse structures are possible, which are ranked by the constraint that the second argument of the condition relation (introduced by *if*) is maximally short: Ideally, this argument should be only C_2 , i.e., the speaker reads a novel if he feels bored, independently of the TV programmes and/or his ability to concentrate. Structures where this argument consists of C_2 and C_3 are less preferred, but still more preferred than structures where this argument is C_2 - C_4 .

Weighted Regular Tree Grammars (wRTGs; Koller et al. 2008) are introduced as a formalism to represent and process partial information on discourse structures in an efficient and expressive way. Preferences as illustrated for (1) are integrated as soft constraints.

References

- Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory. In J. van Kuppevelt and R. Smith, eds., *Current Directions in Discourse and Dialogue*, pages 85–112. Dordrecht: Kluwer.
- Ebert, C. 2005. *Formal investigations of underspecified representations*. Ph.D. thesis, King's College, London.
- Koller, Alexander, Michaela Regneri, and Stefan Thater. 2008. Regular tree grammars as a formalism for scope underspecification. In *Proceedings of ACL-08*. To appear.
- Koller, Alexander and Stefan Thater. 2006. An improved redundancy elimination algorithm for underspecified descriptions. In *Proceedings of COLING-ACL 2006*. Sydney.
- König, Esther and Uwe Reyle. 1999. A general reasoning scheme for underspecified representations. In H. J. Ohlbach and U. Reyle, eds., *Logic, Language and Reasoning. Essays in Honour of Dov Gabbay*, pages 1–28. Dordrecht: Kluwer.
- Reitter, David and Manfred Stede. 2003. Step by step: underspecified markup in incremental rhetorical analysis. In *Proceedings 4th International Workshop on Linguistically Interpreted Corpora (LINC-03)*. Budapest.

REFERENCES / 3

- Stede, Manfred. 2004. The Potsdam Commentary Corpus. In B. Webber and D. Byron, eds., *ACL 2004 Workshop on Discourse Annotation*, pages 96–102. Barcelona, Spain: Association for Computational Linguistics.

