Creating & Analysing Pragmatically Annotated Dialogue Corpora in DART

Full-day practical workshop

Convenor: Martin Weisser

Currently, there is still a dearth of pragmatically annotated corpora. This generally forces researchers in corpus pragmatics to restrict their investigation to more or less fixed and well-known expressions, such as specific discourse markers, different types of requests, or formulaic expressions related to (im)politeness, etc. The Dialogue Annotation and Research Tool (DART) tries to remedy this problem by allowing researchers to build corpora annotated on a number of different pragmatics-relevant linguistic levels easily and efficiently, using a highly readable form of XML. In addition, DART, being a research environment, also provides various options for pre- and post-editing, as well as analysing, corpora compiled in this way.

This workshop aims at allowing interested researchers to familiarise themselves with the architecture and methods employed in DART to construct corpora annotated on multiple pragmatics-relevant levels.

These levels include syntax, semantics, semantico-pragmatics, (surface) polarity, and last, but not least, speech acts, where some 80+ speech-act combinations, including initiation-response sequences, can be annotated fully automatically. The resulting annotations can then be post-edited manually to fix any potential annotation errors, and analysed in various ways discussed as part of the workshop in a gentle step-by-step manner.

The workshop will start with a very brief 'historical' overview of how DART came into existence, but soon move on to discussing the DART format, including conversion from basic transcription formats and other forms of pre-processing the data. This will be followed by an overview of how to create various resources, such as additional lexica, topic files, etc., to customise the annotations, how to run the annotation process itself, as well as conduct any necessary post-processing on the annotated data. And last, but not least, different ways of investigating and analysing the data will be discussed and illustrated.

In order to be able to profit fully from the workshop, participants are also strongly encouraged to bring their own transcribed dialogue data, although most of the initial exercise will rely on sample data provided by me at the beginning of the workshop.

No extensive computing knowledge, other than basic editing experience for linguistic purposes, is required. The software itself is available as freeware under GPL licencing, and can be downloaded from http://martinweisser.org/ling_soft.html#DART. Although it is only available in 32 and 64bit Windows versions, it can also be run through Wine (https://www.winehq.org/) on Mac OS X and Linux. Ideally, participants should download and 'install' the software on their own computers prior to the workshop.