CAT
an advanced environment for the manual annotation of text and corpora
Moretti, Giovanni; Fuoli, Matteo; Sprugnoli, Rachele

2014

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
CAT: an advanced environment for the manual annotation of text and corpora

This software demonstration will provide an overview of the functionality of the Content Annotation Tool – CAT, a general-purpose web-based tool for manual text annotation that can be successfully employed in corpus-based analyses of semantic/pragmatic and discourse phenomena. CAT can also aid the creation of databases of annotated corpus examples for multivariate corpus-based analyses (e.g. Geeraerts et al 1994, Gries 1999, Divjak 2006, Glynn 2009).

Figure 1: the CAT interface

CAT provides a user-friendly interface for annotating text spans of variable length on the basis of an annotation scheme fully defined by the user (Figure 1). Text annotation is performed by highlighting a text span and manually assigning the desired category labels to it. Annotated data can be exported in stand-off XML format or, alternatively, in tabular ‘case-by-variable’ format (Figure 2), which can be used with spreadsheet and statistical software (e.g. R) for further processing and analysis. Finally, the program features a statistics module that calculates the frequency of annotated types and chance-corrected agreement between independent annotators (Dice coefficient).
Among the major strengths of CAT are its ease of use and flexibility. CAT does not require any programming skills or prior knowledge of XML for its installation and use and allows users to freely define and dynamically change the annotation scheme as the project progresses. Compared to similar software, e.g., the UAM Corpus Tool (O’Donnell, 2008), CAT offers several advantages. Most notably, it is web-based, so different people in different locations can work on the same annotation project simultaneously. Further, CAT allows to annotate discontinuous text spans and to export the annotation results in a case-by-variable format, facilitating sophisticated statistical analyses.

CAT has already been used in various Natural Language Processing projects. It has been successfully tested on TimeML annotation for the creation of part of the Ita-TimeBank, the largest Italian corpus annotated with information for temporal processing (Caselli et al., 2011). CAT has also been used to perform a semantic annotation of children’s stories within the TERENCE European project and to manually annotate customer interactions within the EXCITEMENT European project. Recently, CAT has been chosen as the tool for the annotation of temporal information, semantic roles and intra-document co-reference within the NewsReader European project (Fokkens et al., 2013).

While so far it has been mainly used to develop resources for training and evaluation of automatic NLP systems, CAT finds application in the field of corpus linguistics as well as

---

1. [http://www.terenceproject.eu/web/guest/project-overview](http://www.terenceproject.eu/web/guest/project-overview)
2. [http://excitement-project.eu/](http://excitement-project.eu/)
3. [http://www.newsreader-project.eu/](http://www.newsreader-project.eu/)
part of the software demonstration, we will show a concrete example of the use of CAT in a corpus-based multifactorial analysis of evaluation (Bednarek, 2006; Hunston, 2011; Martin and White, 2005) in a small-sized specialized corpus of business reports. The case study will be used to demonstrate the advantages of using manual annotation and CAT for the quantitative analysis of evaluation and to show that insightful multivariate analyses can be performed on the basis of richly annotated corpora.
References


