CAT

an advanced environment for the manual annotation of text and corpora

Moretti, Giovanni; Fuoli, Matteo; Sprugnoli, Rachele

2014

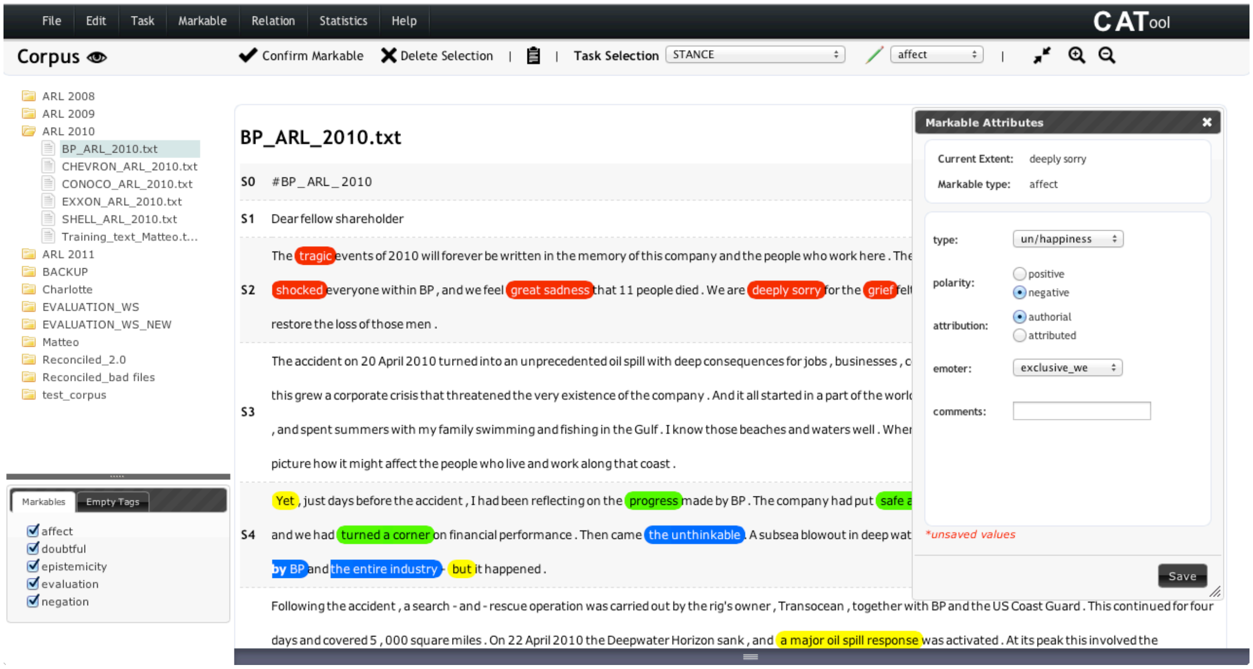
LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

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



BP_ARL_2010.txt

S0 #BP_ARL_2010

S1 Dear fellow shareholder

S2 The tragic events of 2010 will forever be written in the memory of this company and the people who work here. The
shocked everyone within BP, and we feel great sadness that 11 people died. We are deeply sorry for the grief felt
restore the loss of those men.

S3 The accident on 20 April 2010 turned into an unprecedented oil spill with deep consequences for jobs, businesses, c
this grew a corporate crisis that threatened the very existence of the company. And it all started in a part of the worl
, and spent summers with my family swimming and fishing in the Gulf. I know those beaches and waters well. Whe
picture how it might affect the people who live and work along that coast.

S4 Yet, just days before the accident, I had been reflecting on the progress made by BP. The company had put safe a
and we had turned a corner on financial performance. Then came the unthinkable. A subsea blowout in deep wat
by BP and the entire industry - but it happened.

Following the accident, a search - and - rescue operation was carried out by the rig's owner, Transocean, together with BP and the US Coast Guard. This continued for four
days and covered 5,000 square miles. On 22 April 2010 the Deepwater Horizon sank, and a major oil spill response was activated. At its peak this involved the

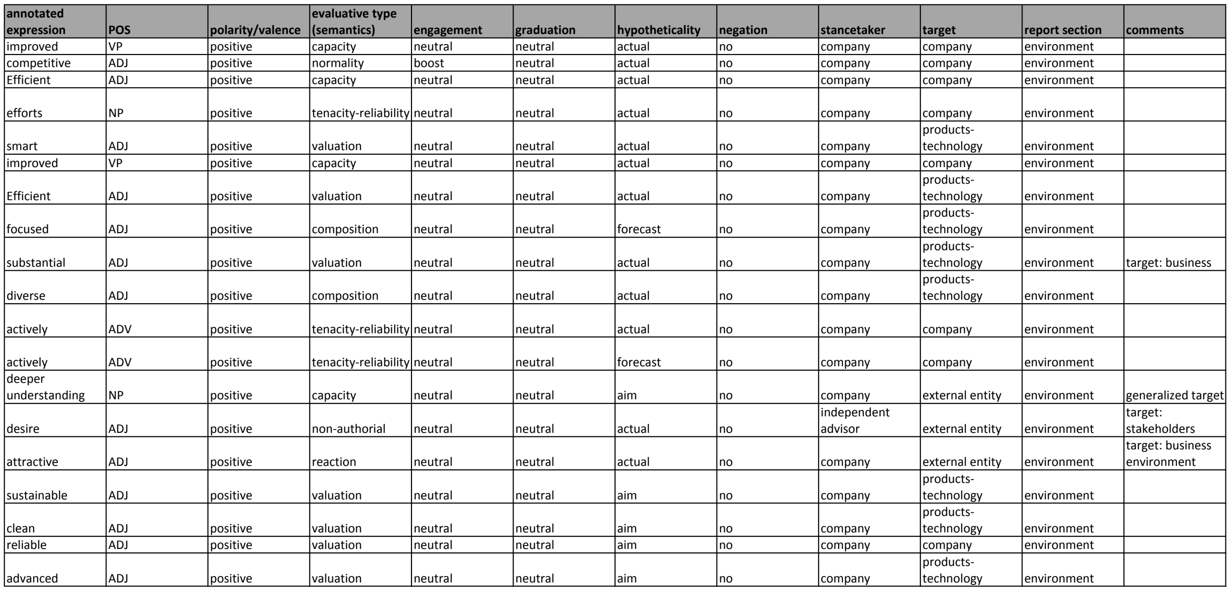
Current Extent: deeply sorry
Markable type: affect

type: un/happiness
polarity: ○ positive ● negative
attribution: ● authorial ○ attributed
emoter: exclusive_we
comments:

*unsaved values

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).

Figure 2: an example of annotated data produced by CAT



annotated expression	POS	polarity/valence	evaluative type (semantics)	engagement	graduation	hypotheticality	negation	stancetaker	target	report section	comments
improved	VP	positive	capacity	neutral	neutral	actual	no	company	company	environment	
competitive	ADJ	positive	normality	boost	neutral	actual	no	company	company	environment	
Efficient	ADJ	positive	capacity	neutral	neutral	actual	no	company	company	environment	
efforts	NP	positive	tenacity-reliability	neutral	neutral	actual	no	company	company	environment	
smart	ADJ	positive	valuation	neutral	neutral	actual	no	company	products-technology	environment	
improved	VP	positive	capacity	neutral	neutral	actual	no	company	company	environment	
Efficient	ADJ	positive	valuation	neutral	neutral	actual	no	company	products-technology	environment	
focused	ADJ	positive	composition	neutral	neutral	forecast	no	company	products-technology	environment	
substantial	ADJ	positive	valuation	neutral	neutral	actual	no	company	products-technology	environment	target: business
diverse	ADJ	positive	composition	neutral	neutral	actual	no	company	products-technology	environment	
actively	ADV	positive	tenacity-reliability	neutral	neutral	actual	no	company	company	environment	
actively	ADV	positive	tenacity-reliability	neutral	neutral	forecast	no	company	company	environment	
deeper understanding	NP	positive	capacity	neutral	neutral	aim	no	company	external entity	environment	generalized target
desire	ADJ	positive	non-authorial	neutral	neutral	actual	no	independent advisor	external entity	environment	target: stakeholders
attractive	ADJ	positive	reaction	neutral	neutral	actual	no	company	external entity	environment	target: business environment
sustainable	ADJ	positive	valuation	neutral	neutral	aim	no	company	products-technology	environment	
clean	ADJ	positive	valuation	neutral	neutral	aim	no	company	products-technology	environment	
reliable	ADJ	positive	valuation	neutral	neutral	aim	no	company	company	environment	
advanced	ADJ	positive	valuation	neutral	neutral	aim	no	company	products-technology	environment	

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[1] and to manually annotate customer interactions within the EXCITEMENT European project[2]. 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[3] (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
[2] http://excitement-project.eu/
[3] 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

Bartalesi Lenzi, V., Moretti, G., Sprugnoli, R., 2012. 'CAT: the CELCT Annotation Tool'. In *Proceedings of LREC 2012*, Istanbul.

Bednarek, M. (2006). *Evaluation in media discourse: analysis of a newspaper corpus*. London & New York: Continuum International Publishing Group.

Caselli, T., Bartalesi Lenzi, V., Sprugnoli, R., Pianta, E., & Prodanof, I. 2011. 'Annotating Events, Temporal Expressions and Relations in Italian: the It-Timeml Experience for the Ita-TimeBank'. In *Proceedings of LAW 5,* 143-151.

Divjak, D. 2006. 'Ways of intending: A corpus-based cognitive linguistic approach to near-synonyms in Russian'. St. Th. & A. Stefanowitsch (eds). *Corpora in Cognitive Linguistics*. Berlin: Mouton de Gruyter, 19–56.

Fokkens A., van Erp, M., Vossen, P., Tonelli, S., van Hage, W., Serafini, L., Sprugnoli, R., Hoeksema, J. 2013. 'GAF: A Grounded Annotation Framework for Events'*. In Proceedings of Workshop on Events: Definition, Detection, Coreference, and Representation*, NAACL HLT 2013, 11–20.

Geeraerts D., Grondelaers, S. & Bakema, P. 1994. *The Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton de Gruyter.

Glynn, D. 2009. 'Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics'. V. Evans & S. Pourcel (eds). *New Directions in Cognitive Linguistics*. Amsterdam: John Benjamins, 77–104.

Gries, Stefan Th. 1999. 'Particle movement: A cognitive and functional approach'. *Cognitive Linguistics* 10: 105-145.

Hunson, S. 2011. *Corpus approaches to evaluation: phraseology and evaluative language*. New York & London: Routledge.

Martin, J. & White, P. 2005. *The language of evaluation: Appraisal in English*. London & New York: Palgrave Macmillan.

O'Donnell, M. 2008. 'Demonstration of the UAM corpus tool for text and image annotation'. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Demo Session*, pages 13–16. Association for Computational Linguistics.