CAT: an advanced environment for the manual annotation of text and corpora

Giovanni Moretti – FBK, Italy
Matteo Fuoli – Lund University, Sweden
Rachele Sprugnoli– FBK/Università di Trento, Italy
Outline

1. Limitations of traditional corpus analysis software (i.e. concordancers)
   – The case of *evaluation* (Hunston and Thompson, 2000; Thompson and Alba-Juez, 2014)

2. Overview of the Content Annotation Tool – CAT

3. Software demonstration
Background
Evaluation

“The expression of the speaker or writer’s attitude or stance towards, viewpoint on, or feelings about the entities or propositions that he or she is talking about” (Hunston and Thompson, 2000, p. 5)
Background
Challenges in the corpus-based analysis of evaluation

1. Open-ended set of forms
2. Multi-word expressions
3. Role of context and co-text
ExxonMobil is dedicated to minimizing adverse risks and impacts associated with our products. (EVALUATIVE)

This may seem strange in a column dedicated to that very subject, but I think it is excellent advice. (NON-EVALUATIVE)
Context/co-text
Evaluative polarity

• Priority issues. Foster a **diverse** work environment that encourages employee growth. (POSITIVE)

• BP operates throughout the world in locations, terrains and climates that are tremendously **diverse** and frequently challenging. (NEUTRAL/NEGATIVE)
Background
Challenges for the quantitative analysis of evaluation

1. It is impossible to identify a definitive finite list of forms that can be searched for using automatic corpus techniques
2. Context needs to be taken into account

‘Top-down’ approach: focus on a restricted range of language forms with predictable evaluative meaning

‘Bottom-up’ approach: manual corpus annotation
The Content Annotation Tool – CAT
The Content Annotation Tool – CAT

Overview

• A general-purpose web-based tool for manual corpus annotation
• User-friendly interface
• Fully customizable annotation scheme
• It allows to annotate text spans of variable length and discontinuous
• It supports multiple annotation layers
• Annotation data stored in stand-off XML format
  – Easily manipulated and converted into tabular ‘case-by-variable’ format
• It features a statistics module
  – Frequency of annotated types and inter-coder agreement
Software demo
The Content Annotation Tool – CAT
Main strengths

• Ease of use and flexibility
• It supports the annotation of discontinuous text spans
• Multiple annotators can access the same project from different locations
• The annotation data are stored in stand-off XML format
  – Flexible and easy to manipulate
  – Easily converted into ‘case-by-variable’ tabular format
  – Supports multiple annotation layers: same tokens and texts can be annotated more than once
The Content Annotation Tool – CAT
Main strengths

• It enables sophisticated statistical analyses based on manual corpus annotation

• It enables **new types of corpus-based analyses**, e.g. quantifying **functions** instead of forms
Evaluation

Variables of interest

• What kind of evaluative meaning is being expressed?
• Who is the stance-taker?
• What/who is being evaluated?
• Are evaluative expressions boosted/hedged?
• What is the topic being discussed?
• What is the discourse genre under analysis?
Fuoli and Glynn (2013)
Coding scheme

14 variables

• Part of speech
• Evaluative semantics (coarse)
• Evaluative semantics (fine)
• Engagement
• Graduation
• Hypotheticality
• Sentential negation
• Target

• Target person
• Stance-taker
• Subject person
• Evaluative polarity
• Topic
• Company
• Year
• Period (before-after)
Fuoli and Glynn (2013)
Statistical analysis

• Univariate statistics
  – Chi-square test

• Exploratory multivariate statistics
  – Correspondence analysis

• Confirmatory multivariate statistics
  – Logistic regression
Multiple correspondence analysis
Evaluative polarity, target, engagement
Fuoli and Glynn (2013)

Conclusions

• Evaluative semantics is not a significant factor
• Stancetaker, hypotheticality and target are the strongest factors
**Fuoli (2013)**

Log-linear analysis of Appraisal in specialized corpus

<table>
<thead>
<tr>
<th>Company</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CHEVRON</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CONOCO</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EXXON</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHELL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th>Markable</th>
<th>Engagement</th>
<th>Judgement</th>
</tr>
</thead>
<tbody>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Standardized Residuals:**

- $\text{< -4}$
- $\text{-4 to -2}$
- $\text{0 to 2}$
- $\text{> 2}$
Accessing CAT

- Beta version can be freely accessed here:
Thank you for listening

Giovanni Moretti (FBK)
moretti@fbk.eu

Rachele Sprugnoli (FBK)
sprugnoli@fbk.eu

Matteo Fuoli (Lund University)
matteo.fuoli@englund.lu.se