# 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

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# 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

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# 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

# Context/co-text

## Polysemy

- 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)

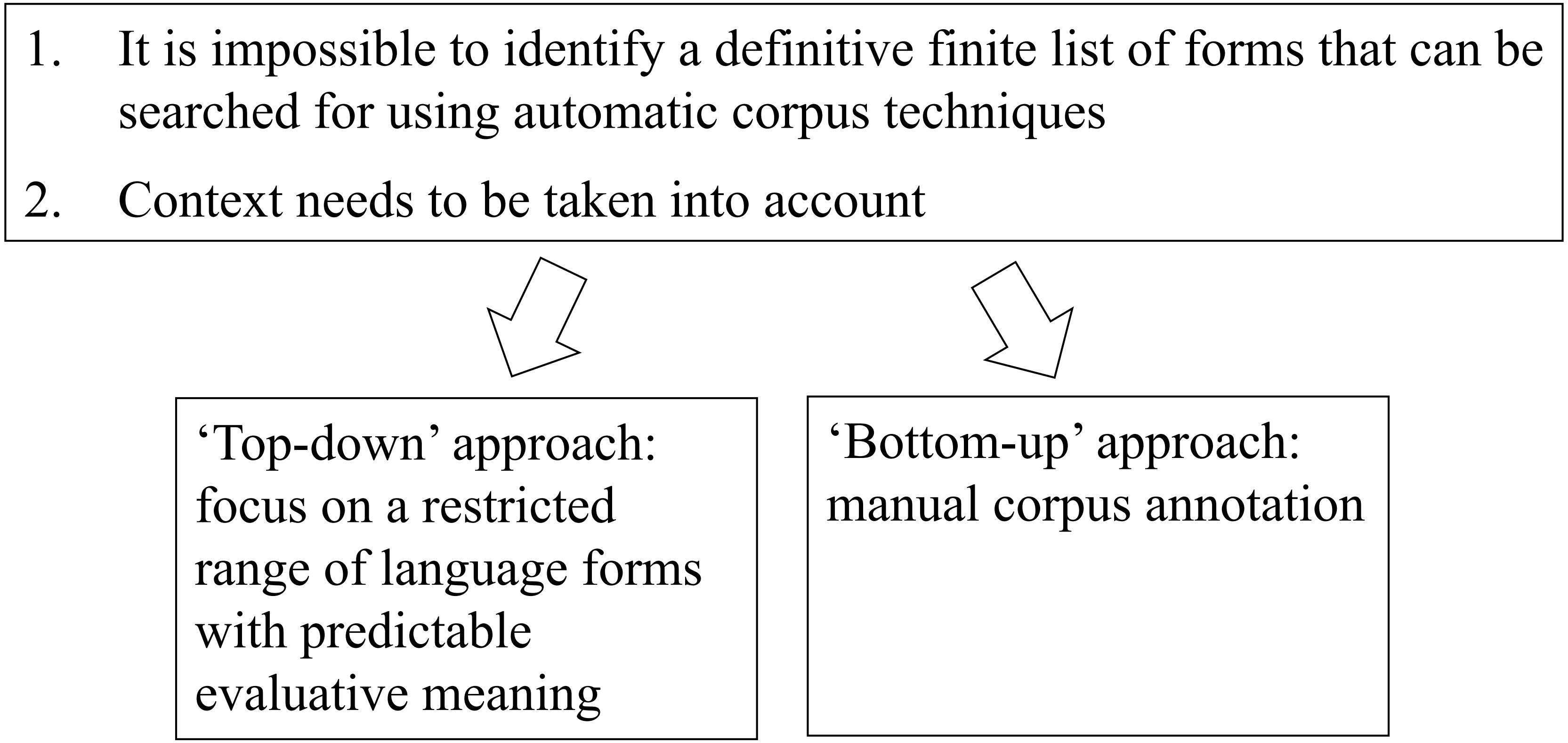
# Context/co-text

## 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

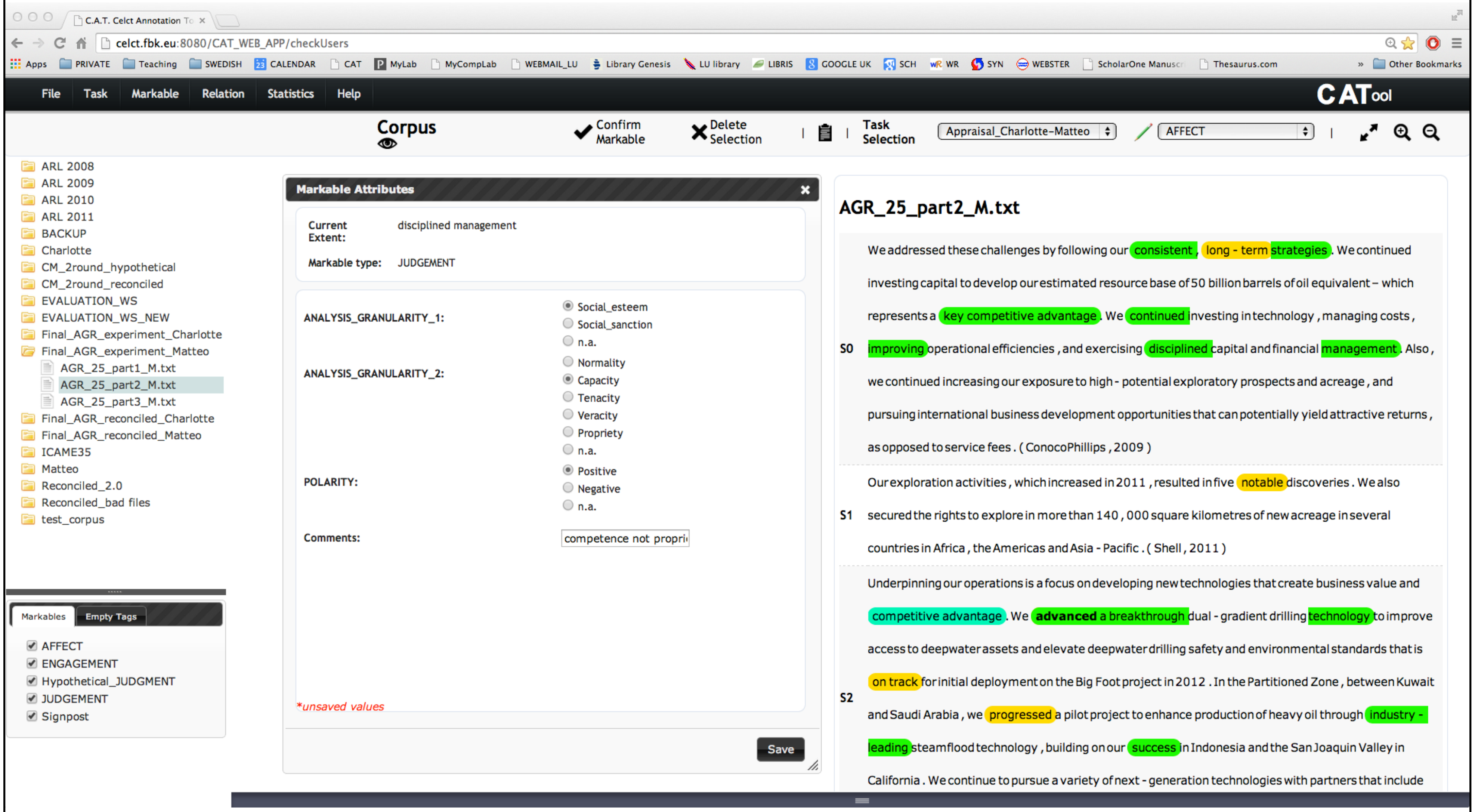
'Top-down' approach: focus on a restricted range of language forms with predictable evaluative meaning

'Bottom-up' approach: manual corpus annotation

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# The Content Annotation Tool – CAT



**Current Extent:** disciplined management

**Markable type:** JUDGEMENT

ANALYSIS_GRANULARITY_1: ● Social_esteem ○ Social_sanction ○ n.a.

ANALYSIS_GRANULARITY_2: ○ Normality ● Capacity ○ Tenacity ○ Veracity ○ Propriety ○ n.a.

POLARITY: ● Positive ○ Negative ○ n.a.

Comments: competence not propri

*unsaved values

## AGR_25_part2_M.txt

S0 We addressed these challenges by following our consistent , long - term strategies . We continued investing capital to develop our estimated resource base of 50 billion barrels of oil equivalent – which represents a key competitive advantage . We continued investing in technology , managing costs , improving operational efficiencies , and exercising disciplined capital and financial management . Also , we continued increasing our exposure to high - potential exploratory prospects and acreage , and pursuing international business development opportunities that can potentially yield attractive returns , as opposed to service fees . ( ConocoPhillips , 2009 )

S1 Our exploration activities , which increased in 2011 , resulted in five notable discoveries . We also secured the rights to explore in more than 140 , 000 square kilometres of new acreage in several countries in Africa , the Americas and Asia - Pacific . ( Shell , 2011 )

S2 Underpinning our operations is a focus on developing new technologies that create business value and competitive advantage . We **advanced** a breakthrough dual - gradient drilling technology to improve access to deepwater assets and elevate deepwater drilling safety and environmental standards that is on track for initial deployment on the Big Foot project in 2012 . In the Partitioned Zone , between Kuwait and Saudi Arabia , we progressed a pilot project to enhance production of heavy oil through industry - leading steamflood technology , building on our success in Indonesia and the San Joaquin Valley in California . We continue to pursue a variety of next - generation technologies with partners that include

# 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

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# Software demo

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# 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

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# 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)

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

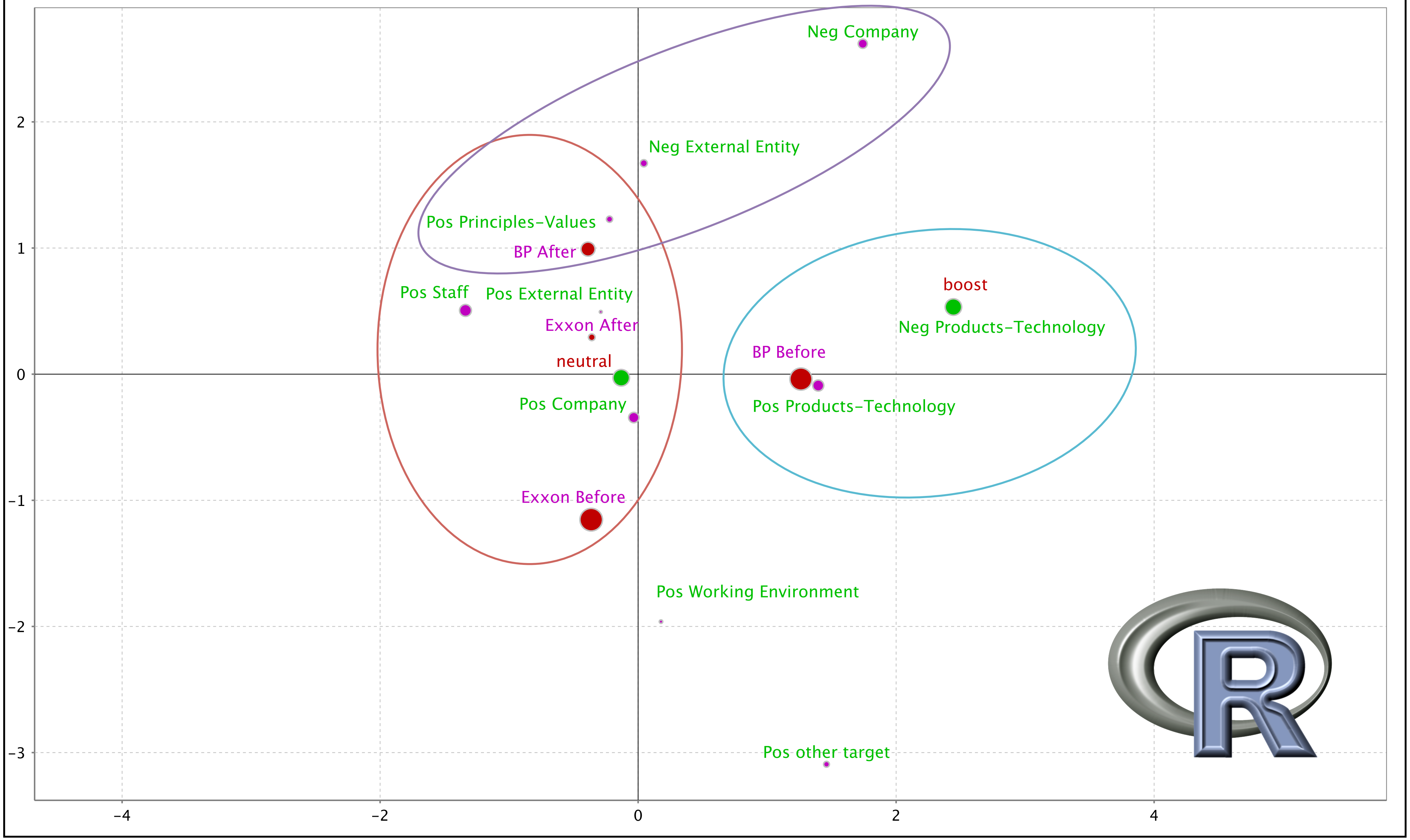
# Fuoli and Glynn (2013)

## 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

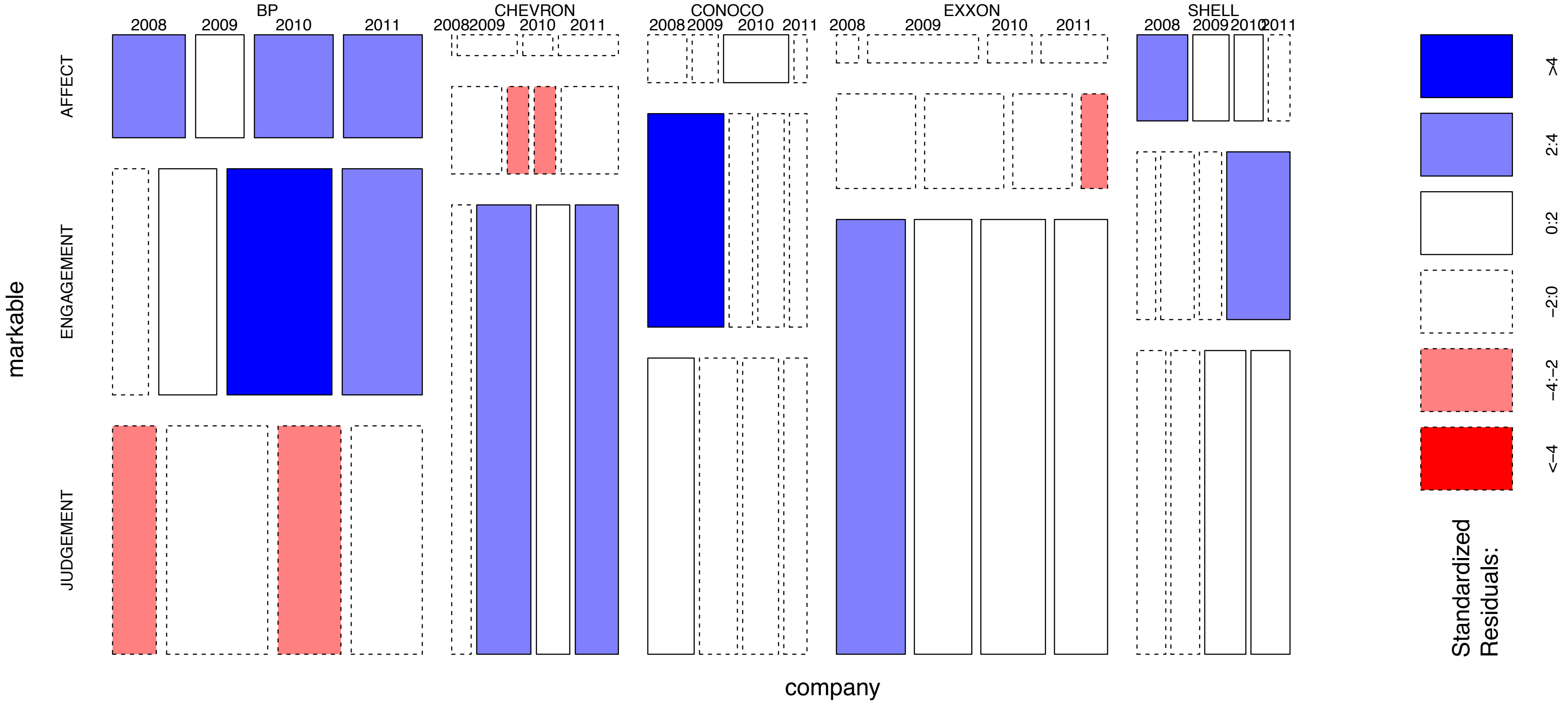
- Evaluative semantics is not a significant factor
- Stancetaker, hypotheticality and target are the strongest factors

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY

# Fuoli (2013)

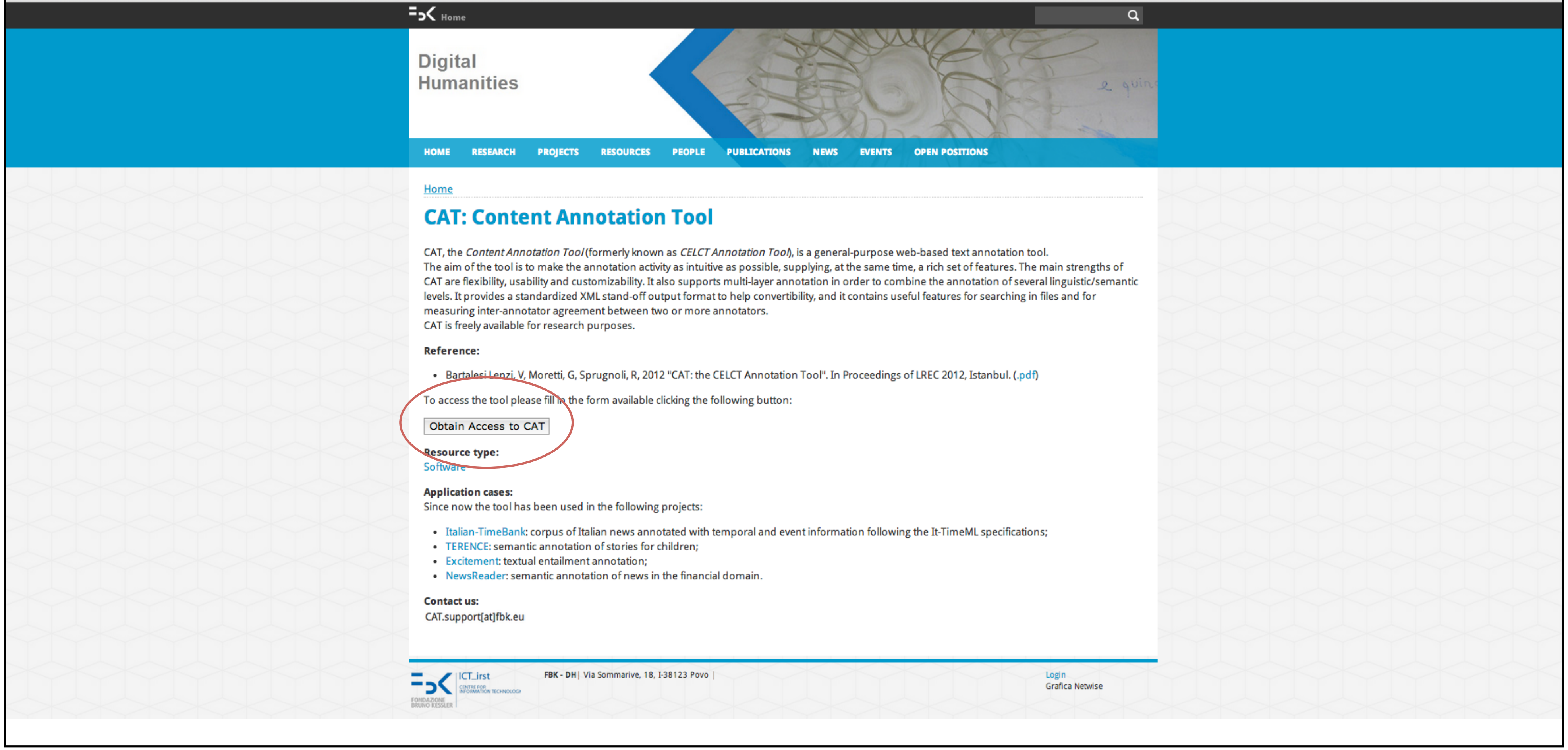
## Log-linear analysis of Appraisal in specialized corpus



# Accessing CAT

- Beta version can be freely accessed here:

  https://dh.fbk.eu/resources/cat-content-annotation-tool



## CAT: Content Annotation Tool

CAT, the *Content Annotation Tool* (formerly known as *CELCT Annotation Tool*), is a general-purpose web-based text annotation tool. The aim of the tool is to make the annotation activity as intuitive as possible, supplying, at the same time, a rich set of features. The main strengths of CAT are flexibility, usability and customizability. It also supports multi-layer annotation in order to combine the annotation of several linguistic/semantic levels. It provides a standardized XML stand-off output format to help convertibility, and it contains useful features for searching in files and for measuring inter-annotator agreement between two or more annotators.
CAT is freely available for research purposes.

**Reference:**

- Bartalesi Lenzi, V, Moretti, G, Sprugnoli, R, 2012 "CAT: the CELCT Annotation Tool". In Proceedings of LREC 2012, Istanbul. (.pdf)

To access the tool please fill in the form available clicking the following button:

**Resource type:**
Software

**Application cases:**
Since now the tool has been used in the following projects:

- Italian-TimeBank: corpus of Italian news annotated with temporal and event information following the It-TimeML specifications;
- TERENCE: semantic annotation of stories for children;
- Excitement: textual entailment annotation;
- NewsReader: semantic annotation of news in the financial domain.

**Contact us:**
CAT.support[at]fbk.eu

ICT_irst
CENTRE FOR INFORMATION TECHNOLOGY
FONDAZIONE BRUNO KESSLER
FBK - DH | Via Sommarive, 18, I-38123 Povo |
Login
Grafica Netwise

# 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

FONDAZIONE
BRUNO KESSLER

LUND UNIVERSITY