Optimising transparency, reliability and replicability: annotation principles and inter-coder agreement in the quantification of evaluative expressions

Fuoli, Matteo; Hommerberg, Charlotte

Published in:
Corpora

DOI:
10.3366/cor.2015.0080

2015

This is a pre-print version of an article that has been published in the journal Corpora, volume 10, issue 3. It can be cited as: Fuoli, M. and C. Hommerberg. 2015. 'Optimizing transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions', Corpora 10 (3), pp. 315–349. For direct quotations and page numbers, please check the published version. The published article can be downloaded from here: http://www.euppublishing.com/toc/cor/10/3

Optimizing transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions

Matteo Fuoli
Lund University, Sweden
matteo.fuoli@englund.lu.se

Charlotte Hommerberg
Linnaeus University, Sweden
charlotte.hommerberg@lnu.se

Abstract

Manual corpus annotation facilitates exhaustive and detailed corpus-based analyses of evaluation that would not be possible with purely automatic techniques. However, manual annotation is a complex and subjective process. Most studies adopting this approach have paid insufficient attention to the methodological challenges involved in manually annotating evaluation, in particular concerning transparency, reliability and replicability. This article illustrates a procedure for annotating evaluative expressions in text that facilitates more transparent, reliable and replicable analyses. The method is demonstrated through a case study analysis of APPRAISAL (Martin and White, 2005) in a small-size specialized corpus of CEO letters published by the British energy company BP and four competitors before and after the Deepwater Horizon oil spill of 2010. Drawing on Fuoli and Paradis' (2014) model of trust-repair discourse, it examines how ATTITUDE and ENGAGEMENT resources are strategically deployed by BP's CEO in the attempt to repair stakeholders' trust after the accident.

Keywords: *evaluation*, APPRAISAL *theory*, *manual corpus annotation*, *inter-coder agreement*, *reliability*, *transparency*, *replicability*, *trust-repair*, *BP Deepwater Horizon oil spill*

1. Introduction

Quantifying evaluation using traditional corpus techniques involves several methodological challenges. Some of the main complexities arise from the fact that

evaluation may be realized by an open-ended set of forms, span multiple words and be expressed both explicitly and implicitly. In addition, it is a highly context-dependent phenomenon. Certain expressions may carry an evaluative meaning in some contexts, but not in others (for an overview, see Hunston, 2011: 12-19).

Traditionally, corpus-based studies have dealt with these challenges by focusing on a restricted range of language forms that tend to have a stable and predictable evaluative meaning (e.g. Biber, 2006; Biber and Finegan, 1989; Camiciottoli, 2013; Conrad and Biber, 2000; Hyland, 1998, 2005; Kaltenbacher, 2006). Within this approach, pre-set lists of attitudinal markers are used to quantify evaluation using automatic techniques. While this method can be useful when dealing with large-scale corpora and when the goal is to identify broad quantitative patterns, it is limited both in terms of coverage, i.e. the range of phenomena that can be accounted for and accurately quantified, and of the level of detail it allows to achieve.

An alternative approach that has been used with increasing frequency (e.g. Bednarek, 2006, 2008; Carretero and Taboada, 2014; Fuoli, 2012; Hommerberg and Don, *forthcoming*; Lipovsky, 2008, 2013; Mackay and Parkinson, 2009; O'Donnell, 2014; Pounds, 2010, 2011; Ryshina-Pankova, 2014; Santamaría-García, 2014; Taboada and Carretero, 2012) is manual corpus annotation. Rather than using pre-defined lists of evaluative expressions, these are manually identified and coded in text, and quantitative data are derived from the annotations. Manual annotation facilitates exhaustive and detailed corpus-based analyses of evaluative language that would not be possible with purely automatic techniques. However, manually annotating evaluation is a complex and subjective task. This may hinder the transparency, reliability and replicability of analyses. With few exceptions (Fuoli, 2012; Hommerberg and Don, *forthcoming*; Read and Carroll, 2010; Ryshina-Pankova, 2014; Taboada and Carretero, 2012), these issues have been largely neglected in the literature. Explicit annotation principles are rarely formulated and, more often than not, no reliability test is performed, resulting in opaque analyses and procedures that are not replicable. Unless these issues are effectively addressed, manual annotation cannot match the robustness of traditional corpus techniques and, as a result, the advantages that it offers are at least partially negated.

This article illustrates a procedure for the manual annotation of evaluative language expressions that is designed to optimize transparency, reliability and replicability. Two strategies are implemented to achieve this goal. First, the annotation guidelines are explicitly formulated and formalized into an annotation manual that is made

publicly available. Second, the reliability and replicability of the annotation procedure are validated by means of an inter-coder agreement test. We demonstrate this method with a case study analysis of APPRAISAL (Martin and White, 2005) in a small-size specialized corpus of CEO letters published by the British energy company BP and four competitors before and after the Deepwater Horizon oil spill of 2010. Drawing on Fuoli and Paradis' (2014) model of trust-repair discourse, the analysis examines how ATTITUDE and ENGAGEMENT resources are strategically deployed by BP's CEO in the attempt to repair stakeholders' trust after the accident.

The primary goal of this article is methodological. It aims to (i) emphasize the importance of transparency, reliability and replicability when subjective interpretations occupy a central role in the analysis, and (ii) offer a concrete illustration of how these aspects can be optimized. The study also contributes to the growing body of literature focusing on BP's communicative response to the Deepwater Horizon oil spill (e.g. Breeze, 2012; Fuoli and Paradis, 2014; Harlow et al., 2011; Muralidharan et al., 2011; O'Connor, 2011; Schultz et al., 2012; Wickman, 2013), by investigating the company's post-crisis discourse from the point of view of trust-repair.

The article is organized as follows. Sections 2 and 3 provide the background for the study, outlining Fuoli and Paradis' (2014) model of trust-repair discourse and APPRAISAL theory. Section 4 describes the corpus. Section 5 presents the method, including a description and the results of the inter-coder agreement test. Section 6 presents the analysis. Section 7 concludes by assessing the method and offering suggestions for future work.

2. Trust-repair discourse

Trust is a valuable asset for a company. A high level of trust may, among other things, facilitate business transactions, increase employee commitment and customer satisfaction, and help improving relations between the organization and its stakeholders, thereby lowering the risk of uncooperative behavior and conflict (Barney and Hansen, 1994; García-Marzá, 2005; Ingenhoff and Sommer, 2010; Pirson and Malhotra, 2011). More generally, trust is key to ensuring a company's social legitimacy and, ultimately, its long-term viability (Poppo and Schepker, 2010: 124).

Trust is a dynamic and constantly changing construct (Cook, 2001; Linell and Marková, 2013; Marková and Gillespie, 2008). Episodes of wrongdoing or negligence, controversies and scandals negatively affect stakeholders' trust in a company. A recent and illustrative example of this is BP's Deepwater Horizon oil spill. The Deepwater Horizon oil spill is considered the largest accidental marine oil spill in history (Robertson and Krauss, 2010). It was triggered by the explosion and sinking of an offshore oil rig controlled by the British energy company BP in the Gulf of Mexico, on the 20th of April 2010. The explosion resulted in eleven fatalities and the oil spill that followed has inflicted severe damage to both the environment and the economy of the Gulf. The accident has severely impacted BP's image and reputation. In particular, the controversies surrounding BP's response to the spill undermined its trustworthiness, plunging the company into a crisis of public trust.

During a corporate crisis, discourse acquires a central role in controlling and minimizing the damage to a company's image (Benoit, 1997). Fuoli and Paradis (2014) present a theoretical framework for investigating and explaining trust-repair discourse. The authors propose that, when trust is at stake, the trust-breaker will shape their discourse to try to influence the trust-giver's impressions of their *trustworthiness*. Trustworthiness is composed of three main components: *ability*, *integrity* and *benevolence* (Mayer et al., 1995). Ability concerns how competent and skillful a person (or company) is, integrity relates to how honest and sincere they are, and benevolence refers to the extent to which a person is believed to care about the trust-giver, beyond self-interested concerns. When a trust-breaking event occurs, the trust-breaker's trustworthiness will be at stake along one or more of these dimensions. Their trust-repair discourse will be geared towards reshaping the trust-giver's impressions accordingly.

Fuoli and Paradis (2014) identify two fundamental strategies that the trust-breaker may pursue to achieve this communicative goal. On the one hand, the trust-breaker may foreground their goodwill, sympathy and positive qualities, a strategy which the authors name *emphasize the positive*. On the other hand, the trust-breaker may seek to dialogically engage with and act upon the discourses that generate distrust. The authors term this strategy *neutralize the negative*. Both strategies may be used simultaneously and interact in a single instance of trust-repair discourse.

Fuoli and Paradis (2014) propose that the emphasize-the-positive and neutralize-the-negative strategies are realized in text through attitudinal and dialogic engagement resources, respectively. Affective language may be used, for example, to

communicate empathy and solidarity, promoting a positive impression of the trust-breaker's benevolence, as in (1).

(1) We are deeply sorry for the grief felt by their families and friends. We know nothing can restore the loss of those men.

Evaluative expressions may be strategically deployed, for instance, to emphasize the trust-breaker's ability, as in (2).

(2) We recognize there is a great deal more to do, but I can report that BP finished its year of consolidation in robust shape.

Dialogic engagement concerns the linguistic devices by which speakers mark their stance towards other opinions and includes e.g. negation/denial, adversative discourse markers, epistemic modals (Martin and White, 2005; White, 2003). These resources may be used to respond to and seek to neutralize the discourses that represent a source of distrust, as the following example from the corpus shows.

(3) Our fundamental purpose is to create value for shareholders, but we also see ourselves as part of society, not apart from it.

In (3), through the use of the contrastive and negation markers *but* and *not*, BP's CEO engages with and strongly rejects the view that the company is solely focused on protecting shareholders' interests and is insensitive to society's legitimate concerns after the spill. By doing this, he seeks to repair the company's integrity.

In the analysis presented in section 6, we use Fuoli and Paradis' (2014) model to investigate the trust-repair discourse strategies employed by BP's CEO in his letters to shareholders after the Deepwater Horizon oil spill. The first step in the analysis consists in identifying and analyzing the use of attitudinal and dialogic engagement expressions, which, as discussed above, are predicted by the model to play a crucial communicative role in trust-repair discourse. To accomplish this task, we employ APPRAISAL theory (Martin and White, 2005), which provides an integrated and coherent framework for the analysis of these discursive phenomena. The next section outlines the theory.

3. APPRAISAL theory

APPRAISAL theory offers a framework for analyzing how speakers negotiate and maneuver interpersonal roles and relations in discourse. The type of linguistic phenomena accounted for by the theory have also been investigated under different labels from alternative and partly overlapping perspectives, e.g. *evaluation* (Hunston, 2011; Hunston and Thompson, 2000; Thompson and Alba-Juez, 2014) and *stance* (Biber, 2006; Biber and Finegan, 1988, 1989; Conrad and Biber, 2000).

Grounded in Systemic Functional Linguistics (Eggins, 2004; Halliday, 1994), the APPRAISAL framework is presented as a set of choices that are available to the speaker for expressing and negotiating interpersonal stances and value positions in discourse. The model includes three interactive components: ATTITUDE, ENGAGEMENT and GRADUATION. ATTITUDE concerns feelings, such as emotional reactions, judgments of behavior and evaluations of things. ENGAGEMENT comprises a set of resources by means of which speakers adopt a position with respect to alternative opinions and voices. GRADUATION is used for scaling the intensity of an ATTITUDE or the degree of speaker investment in a proposition (Martin and White, 2005: 35-39). These three categories are further developed into an extensive range of subcategories organized in networks.

The analysis presented below is limited in scope to the ATTITUDE subcategories of AFFECT and JUDGEMENT, and ENGAGEMENT. The analysis of JUDGEMENT is further restricted to instances of '*self*-JUDGEMENT', i.e. expressions that are used by the CEOs to evaluate their company members' qualities and behavior (see section 5.2.1.). Based on Fuoli and Paradis' (2014) model, these resources can be expected to play a key role in trust-repair discourse, and their frequency can thus be anticipated to significantly change in BP's CEO letters after the spill. In section 6 we explore this hypothesis. While other APPRAISAL resources may also play a role in trust-repair discourse, we decided to focus on those that, based on Fuoli and Paradis' (2014) model, can be expected to be most directly relevant and thus provide the most meaningful insights into BP's trust-repair strategies. In light of these restrictions, the following sections provide an overview of AFFECT, JUDGEMENT and ENGAGEMENT only. A complete description of the APPRAISAL model can be found in Martin and White (2005).

3.1 ATTITUDE

ATTITUDE refers to the expression of feelings and evaluations in discourse. The APPRAISAL model suggests a division of this area of interpersonal meaning into three regions: AFFECT, JUDGEMENT and APPRECIATION. As mentioned above, APPRECIATION was not included in the analysis, and will thus not be considered in detail here.

AFFECT concerns emotions and states of mind. It can be realized by a variety of linguistic expressions belonging to any word class, e.g. adjectives (e.g. *happy*/*sad*), nouns (e.g. *joy*/*sorrow*), verbs (e.g. *love*/*hate*) or adverbials (e.g. *happily*/*sadly*). Expressions of AFFECT may have positive or negative *valence*, i.e. they may be used to communicate positive or negative emotions. The APPRAISAL model classifies AFFECT into four main types according to semantic/functional criteria: DIS/INCLINATION, UN/HAPPINESS, IN/SECURITY, DIS/SATISFACTION (Martin and White, 2005: 71). Bednarek (2008: 172) modifies the AFFECT subsystem by creating an independent category for SURPRISE (previously included in IN/SECURITY). The following is an example of AFFECT from our corpus.

(4) The explosion and fire on the Deepwater Horizon rig shocked everyone within BP and we feel great sadness that 11 people died.

JUDGEMENT refers to positive and negative evaluations of people according to different parameters. The system consists of two main subcategories: SOCIAL SANCTION, which refers to the moral evaluation of people's behavior (veracity and propriety), and SOCIAL ESTEEM, which concerns their normality, capacity or tenacity. Similarly to AFFECT, JUDGEMENT can be instantiated by an open-ended set of linguistic expressions, including adjectives (e.g. *skilled*, *clever*, *courageous*), manner adverbials (e.g. *cleverly*, *in a diligent and responsible way*) and nouns (e.g. *success*, *leader*). (5) shows an example of JUDGEMENT from the corpus.

(5) With the talent and commitment of the people of ExxonMobil, we are strong, resilient and well-positioned for the future.

All the examples given so far account for what Martin and White (2005) term *inscribed* ATTITUDE, i.e. feelings and evaluations that are explicitly conveyed by manifestly positive or negative wordings. As the authors note, however, in certain contexts 'the selection of ideational meanings [may be] enough to invoke evaluation, even in the absence of attitudinal lexis that tells us directly how to feel' (Martin and

White, 2005: 62). Thus, in a persuasive text such as a CEO letter, and in a situation where the trustworthiness of the company has been damaged, a factual statement like (6) may be interpreted as implying positive evaluation, even though no explicit assessment is expressed.

(6) To encourage excellence in risk management throughout the organization, we are reviewing how we incentivize and reward people.

Martin and White (2005) term this type of implicit evaluations *invoked* ATTITUDE. While invoked realizations may be relevant to understanding the discursive dynamics of trust, it is very hard, if at all possible, to reliably identify and quantify instances of this type. The present study is therefore restricted to inscribed ATTITUDE (see section 5.2.1).

3.2 ENGAGEMENT

The APPRAISAL system of ENGAGEMENT comprises linguistic resources used by speakers to indicate their stance towards alternative opinions and viewpoints, and to anticipate and manage potential responses from interlocutors (White, 2012). ENGAGEMENT incorporates a wide range of diverse phenomena that have traditionally been referred to in the linguistics literature with labels such as *modality*, *polarity*, *evidentiality* and *attribution* (Martin and White, 2005: 94). The function of these resources in discourse is examined from a dialogic perspective (Bakhtin, 1981); they are used by speakers to signal whether they anticipate their proposition to be controversial or likely to be questioned by the audience, and maneuver potential reactions and responses.

ENGAGEMENT resources are organized into different categories based on their communicative/dialogic function. A major dividing line is drawn between *monoglossic* and *heteroglossic* utterances. Monoglossic propositions are those in which other viewpoints are not recognized, i.e. bare or categorical assertions. Heteroglossic resources are broadly subdivided into those that entail dialogic *expansion* and those that involve dialogic *contraction* (Martin and White, 2005: 102). Dialogic expansion incorporates resources by which the dialogic space is opened up for alternative viewpoints and voices, while dialogic contraction subsumes options that serve the communicative purpose of challenging or restricting the scope for

alternative positions and voices.

The ENGAGEMENT category of dialogic expansion involves two subcategories. ENTERTAIN refers to options that signal that the position advanced is to be seen as just one voice among others on a particular issue, i.e. to indicate that the speaker takes into consideration the possible existence of alternative viewpoints in addition to the one they are advancing. The resources that are subsumed under this heading include modal auxiliaries (*may*, *might*, *could*), modal adverbs (*perhaps*, *probably*), epistemic mental predicates (*think*, *suspect*, *doubt*) and certain evidentials (*it seems*, *it appears*) (Martin and White, 2005: 105). The following is an example of ENTERTAIN from the corpus.

(7) Despite these actions, ConocoPhillips considers it possible that the recession could restrain energy demand and prices for several years.

The other dialogic expansion subcategory, ATTRIBUTION, refers to linguistic resources by which the proposition is attributed to an external source, and speakers present themselves as having no stake in it, i.e. as simply conveying information. This category may be realized, for example, by reporting structures (*x claims*, *believes*, *suggests*), nominalizations of such structures (*assertion that*, *claim that*) or adverbial adjuncts (*according to*). An example of ATTRIBUTION from the corpus is the following.

(8) The report stated that decisions made by multiple companies and work teams contributed to the accident.

In contrast to communicative strategies of dialogic expansion, which serve the purpose of opening up the dialogic space for alternative opinions, dialogic contraction operates to exclude these alternatives. The resources of dialogic contraction are subdivided into two broad categories. The category of DISCLAIM subsumes formulations that invoke alternative viewpoints only to directly reject or replace them (Martin and White, 2005: 118). DISCLAIM encompasses negation/denial, adversative discourse markers (e.g. *however*, *but*, *yet*) and other wordings by which speakers indicate that the natural expectations arising from a proposition are not fulfilled (e.g. *surprisingly*, *even though*) (Martin and White, 2005: 120-121). An example of this category is provided below.

(9) Our investigation report was published on 8 September 2010, and found that

no single factor caused the accident.

The other main subgroup of dialogic contraction is PROCLAIM. This category includes expressions which, rather than directly rejecting other viewpoints, act to limit the scope for alternative positions (Martin and White, 2005: 121). Wordings that are classified under this category include those that signal agreement between the speaker and the addressee and thereby exclude alternative voices (e.g. *of course*, *admittedly*, *obviously*), wordings whereby the presence of the authorial voice is emphasized so as to suppress any resistance that might exist (e.g. *I contend*, *indeed*, *we firmly believe that*) and expressions of endorsement of attributed propositions (e.g. *the report found that*, *the studies demonstrate that*). (10) is an example of PROCLAIM from the corpus.

(10) I want to make it absolutely clear that we are not seeking a return to business as usual.

The categories outlined in this section were included in the corpus annotation scheme, which is described in section 5.2.1. The next section describes the corpus.

4. Data

For this study we compiled a specialized corpus of CEO letters published by BP and four other major oil companies during a period of two years before and after the Deepwater Horizon oil spill. In section 6.1, the BP letters published after the spill are compared with those published before the accident and by the other companies with the aim to identify significant patterns of change in BP's discourse and gain insights into the company's trust-repair strategies. The corpus is presented in section 4.2. The next section briefly describes the genre of CEO letters.

4.1 The genre of CEO letters

CEO letters are an important genre of corporate public discourse, being 'the most prominent and widely read part of an annual report' (Hyland, 1998: 224). From a communicative point of view, these letters play a fundamental role in framing the information included in the reports, steering the reader's response to the facts proffered, and promoting a positive corporate image (Garzone, 2005; Hyland, 1998).

CEO letters can be considered a promotional genre and their rhetorical purpose is primarily persuasive; they aim to convince stockholders of the validity of their investment in a company and seeking continued financial support at the same time (Breeze, 2012, 2013; Hyland, 1998).

Unlike the more technical parts of the annual report, which include audited information about a company's financial performance, CEO letters are not subject to any legal obligations to truthfulness and transparency (Breeze, 2012: 6), but rather represent the CEO's own perspective on the company's performance. CEO letters thus usually exhibit a more personalized and overtly evaluative writing style, which is manifested in structural features typical of the letter genre, such as a salutation or direct address and the 'real' CEO's signature, the frequent use of personal pronouns and evaluative expressions to express the speaker's stance (Bhatia, 2004; Garzone, 2005; Gillaerts and Van de Velde, 2011; Hyland, 1998). Compared to the rest of the annual report, whose chief function is to provide detailed information about a company's financial standing, CEO letters can be seen to primarily perform an interpersonal function, putting a face to the company and establishing a dialogue with the readership. Accordingly, these letters have 'enormous rhetorical importance in building credibility and imparting confidence, convincing investors that the company is pursuing an effective strategy' (Hyland, 1998: 224).

4.2 The corpus

Table 1 provides a detailed breakdown of the corpus contents.

Table 1. Corpus details

The corpus includes 20 CEO letters[1], of an average length of 1192 words. The total corpus size is 23484 tokens. The letters are included in the companies' annual reports, which are public documents that can be freely downloaded from their websites.

5. Methods

The analysis presented in section 6 combines quantitative and qualitative methods.

Quantitative analysis is used to (i) identify patterns of change in BP's use of AFFECT, JUDGEMENT and ENGAGEMENT resources and (ii) guide the qualitative analysis of the BP letters published after the spill. The qualitative analysis consists in an interpretive close reading of BP's letters after the spill, which aims to elucidate the rhetorical function of the annotated features and investigate their interplay with other discursive resources and motives. The results of both analyses are interpreted in light of Fuoli and Paradis' (2014) model of trust-repair discourse, in order to seek to explain the patterns observed and determine the strategies deployed by the CEO to repair stakeholders' trust after the spill. The quantitative data are derived from the manual annotation of the corpus. This section describes the corpus annotation process and the inter-coder agreement test that was conducted to validate the reliability and replicability of the annotations. This is preceded by a brief discussion of the limitations of automatic corpus methods for the analysis of evaluation.

5.1 Limitations of automatic corpus methods for the analysis of evaluation

Quantifying evaluative expressions using automatic corpus methods involves several challenges. Some of the main complexities are due to well-known properties of evaluation (for a review, see Hunston, 2011: 12-19). In particular:

- Evaluation may be realized by an open-ended range of expressions. Thus, analyses based on pre-set lists of evaluative forms will not, in most cases, be exhaustive, i.e. they will not cover *all* evaluation in a text or corpus. This also applies to cases where pre-defined sets of lexico-grammatical patterns are used as a starting point (e.g. Bednarek, 2009; Hunston and Sinclair, 2000).
- Context is crucial for decoding evaluative meaning. Context and co-text play a critical role in determining the evaluative or non-evaluative nature of an expression, and thus need to be taken into account to correctly identify relevant forms. This applies to both explicit and implicit evaluation. Compare, for example:

(11) a. ExxonMobil is dedicated to minimizing adverse risks and impacts associated with our products. [Explicitly evaluative: 'devoted to a cause, ideal, or purpose']

b. This may seem strange in a column dedicated to that very subject,

but I think it is excellent advice. [Non-evaluative: 'used for one particular purpose']

This also holds true for very frequent polysemous ENGAGEMENT markers such as *know*, *show* or *understand*, which perform an epistemic/dialogic function in certain co-texts only (Fuoli, 2012: 66). Classifying evaluative expressions is also often impossible without taking co-text into account. For instance, the adjective *great* in (12) can only be classified as an instance of JUDGEMENT based on the co-textual environment in which it appears, rather than on inherent semantic aspects of the word.

(12) It has truly been an honour for me to serve 12 years as chairman of what is one of the world's great enterprises.

Therefore, corpus techniques that treat words and expressions in isolation from their original co-texts, e.g. word-list or n-gram analysis, will make accurate analysis difficult to achieve.[2]

- Evaluative items may span multiple words (e.g. *long-term commitment to excellence*). Accordingly, automatic methods based on single words or n-grams may fail to recognize their 'real' boundaries. This might affect accuracy. If wordlists were used as a starting point to quantify evaluation, for instance, *commitment* and *excellence* in the example above would be counted as two evaluative items, where we believe they are better analyzed as part of one single evaluative expression instead. Similarly, the words *considers* and *possible* in (7) might erroneously be counted as two independent instances.

Automatic semantic tagging systems are, to date, insufficiently accurate, and often produce false positives and coding errors (see e.g. Murphy, 2013). Therefore, they do not offer a reliable alternative to traditional corpus methods. Conversely, manual corpus annotation allows to overcome the challenges described above. By manually annotating text, all evaluative expressions can be identified and counted.[3] Identification and classification can be more accurate, as context and co-text are properly accounted for. However, manual corpus annotation is a complex and subjective process. This poses challenges to achieving transparent, reliable and replicable analyses. The next sections describe the strategies adopted in this study to address these issues.

5.2 Annotation procedure

The annotation of our corpus comprised the following main steps:

1. Annotation scheme design. As the first step in the process, we needed to establish principles for adapting the general-purpose APPRAISAL framework to the goals of our analysis and the specificities of the texts under study (Hommerberg and Don, *forthcoming*), and create context-specific guidelines to implement the framework to the annotation of our corpus. This phase involved decisions concerning the level of granularity to adopt in the analysis, what context-specific guidelines to follow when identifying and classifying instances, and what rules to apply when encountering ambiguous items. The final product of this phase was a context-specific annotation manual, which is reported in the Appendix. The annotation scheme is presented in section 5.2.1.

2. Annotator training and inter-coder agreement test. Manually annotating APPRAISAL is a complex and subjective process. This may affect the reliability of the corpus annotation and of the quantitative data derived from it. In addition, it may hinder the replicability of the annotation procedure. Different people may produce different analyses, unless transparent and shared annotation principles are established. The first measure we took to improve these aspects was that of creating a manual that contains explicit annotation guidelines for all annotators to follow. In addition, the annotation of the whole corpus was preceded by an inter-coder agreement test. A detailed account of the test is provided in section 5.3. One of the outcomes of the test was a jointly annotated 'gold standard' sample that was used as a guide for the annotation of the whole corpus.

Fig. 1. The CAT user interface

3. Corpus annotation. Author 1 manually annotated and classified all instances of the categories included in the annotation scheme (Fig. 3). The annotation was performed based on the guidelines outlined in the annotation manual and the gold-standard sample created as a result of the inter-coder agreement test. Examples of APPRAISAL expressions included in Martin and White (2005) and Bednarek (2008) were also consulted during this task. Every text in the corpus was examined at least twice, at

different points in time. The annotation required around 25 hours of work, i.e. approximately 1 hour per thousand words. The software used for this task is the *Content Annotation Tool* or 'CAT' (Bartalesi Lenzi et al., 2012).[4] CAT is a web-based annotation tool that provides a user-friendly interface for annotating words and expressions in text (see Fig. 1). Coding schemes can be fully customized to fit the annotation task at hand.

Fig. 2. The CAT output

CAT stores the annotations as stand-off XML. This format can easily be converted into a tabular *case-by-variable* format (see Fig. 2), which can be used with spreadsheet and statistical software for further processing and analysis.

5.2.1 Annotation scheme

Fig. 4 shows the scheme used for the annotation of the corpus.

Fig. 3. Annotation scheme[5]

As mentioned above, the analysis was restricted to the ATTITUDE subsystems of AFFECT and JUDGEMENT, and ENGAGEMENT. The annotation of JUDGEMENT was restricted to (i) positive instances and (ii) evaluative expressions that are used to praise the company and its members, excluding other evaluative 'targets'. In this way, we could isolate those evaluative expressions that may directly realize the emphasize-the-positive strategy, in accordance with Fuoli and Paradis' (2014) model. In addition, only explicit (inscribed) instances of APPRAISAL were annotated. Invoked instances were found to be extremely difficult to reliably identify and were, therefore, ignored in the annotation process. Further, a distinction was made between evaluative expressions that are included in *realis* contexts, and those that are included in *irrealis*, non-factual contexts, i.e. expressions that are not used to afford an actual evaluation

of the company but rather to refer to future or hypothetical desirable states of affair (Taboada et al., 2011: 278-279). Compare, for example:

(13) a. I would also like to thank our people for the tremendous effort, dedication and passion they have shown.

b. We can never eliminate every hazard, but we can become an industry leader in limiting and understanding risk.

While both types of expressions may play a role in trust management, irrealis evaluations were eventually excluded from the analysis, on the grounds that robust agreement on their identification could not be reached. More detailed information about coding decisions can be found in the annotation manual (Appendix).

5.3 Inter-coder agreement test

In the context of a corpus annotation task, *inter-coder agreement* is a measure of the extent to which independent annotators make the same decisions when assigning pre-defined categories to units of text. Inter-coder agreement can be used as a criterion for evaluating the *reliability* of the corpus annotation, based on the assumption that 'data are reliable if coders can be shown to agree on the categories assigned to units to an extent determined by the purposes of the study' (Artstein and Poesio, 2008: 557). Demonstrating reliability is necessary to ensure the *reproducibility* of an analysis (Krippendorff, 2004: 215). If independent annotators using the guidelines given consistently make equivalent coding choices, we can expect other annotators to produce similar interpretations, given similar circumstances. Conversely, a low level of agreement may indicate either that the coding scheme is defective or not sufficiently explicit, or that the annotators need more training. Reliability is a pre-requisite for the *validity* of a coding scheme, i.e. the extent to which it captures the 'truth' about the phenomenon under investigation (Artstein and Poesio, 2008; Krippendorff, 2004). As Krippendorff (2004: 213) remarks, however, 'reliability is a necessary, but not a sufficient, condition for validity'. Independent analysts may, in fact, share similar interests, biases and prejudices that may affect the way they interpret the data in similar ways.

Inter-coder agreement measures have been widely used in Computational Linguistics, where a growing need for manually annotated corpora in the development of

computational models of semantic and pragmatic phenomena has been accompanied by concerns about the subjectivity involved in the analysis of these aspects of language (Artstein and Poesio, 2008). Despite the importance of manual annotations in the creation of linguistic corpora, measures of inter-coder agreement are still rarely reported in the corpus linguistics literature (Spooren and Degand, 2010; Voormann and Gut, 2008). This also applies to research using the APPRAISAL framework where, in spite of criticisms concerning the arbitrariness of its taxonomy and the subjectivity of the analyses (Taboada and Carretero, 2012: 278), explicit accounts of inter-coder agreement are not commonly provided. There is, however, a growing awareness of the importance of accounting for reliability in this area, and several recent studies based on APPRAISAL theory include a discussion of inter-coder agreement (Fuoli, 2012; Hommerberg and Don, *forthcoming*; Read and Carroll, 2010; Ryshina-Pankova, 2014; Taboada and Carretero, 2012).

Annotating APPRAISAL can be a very complex and subjective task. This is mainly due to the fact that the framework itself is conceived as a flexible interpretive tool, a 'basic draft of categories' (Hommerberg and Don, *forthcoming*), rather than a definitive model of evaluation that can be applied to any kind of text in a mechanical way. These complexities are heightened in the present context by the fact that the very communicative goal of the texts considered is that of promoting a positive image of the companies. This implies that all utterances included in the texts may be safely interpreted as conveying evaluation, which makes it very hard to confidently identify and annotate units of 'explicit' APPRAISAL. As a result, reaching high inter-coder agreement scores may be, in this context, a particularly challenging task, and the ideal of coders who 'work completely independently and agree substantially' (Spooren and Degand, 2010: 253) difficult to reach. According to Spooren and Degand (2010), there are three main strategies that can be applied in situations where reaching high inter-coder agreement between independent coders is challenging:

1. Double coding. Two annotators code the entire data independently and then discuss all the disagreements until full consensus is reached. This strategy improves the quality of the annotation as it promotes cooperative rather than idiosyncratic coding strategies, and because the coders are forced to make their reasoning explicit and convince each other in case of disagreement. This is, however, a very time consuming procedure.

2. Partial overlap between two or more coders. A portion of the data is double coded, while the rest of the corpus is annotated by only one person. Inter-

coder agreement is calculated on the double-coded sample. Disagreements are discussed and reconciled between the annotators to enhance the quality of the single-coded data.

3. One coder does all. The entire corpus is annotated by only one coder. While the annotator can be expected to adopt subjective annotation strategies, we can assume these strategies to be systematic. This means that if the annotator has a tendency to over-annotate one specific category, that category will be globally overrepresented in the analysis, which should not represent an obstacle to answering the research question.

Due to time and resource constraints, double coding the entire corpus was not a viable option. The one-coder-does-all alternative was also rejected, on the grounds that it is the weakest form of reliability check. Given the nature of the task at hand, it would leave the possibility open for individual biases to substantially affect the outcome of the analysis. Therefore, for this study we have adopted the second of the above solutions.

5.3.1 Inter-coder agreement test design

The design of our inter-coder agreement test closely resembles that of Read and Carroll (2010). The test involved two expert annotators, i.e. author 1 and author 2. Both annotators had previous experience with corpus annotation using the APPRAISAL framework. The test consisted of two related annotation tasks:

1. Identification of expressions of APPRAISAL, a process which can be referred to as *markable identification* (Artstein and Poesio, 2008) or *unitization* (Krippendorff, 2004);

2. Classification of the expressions identified according to the scheme outlined in section 5.2.1.

These tasks were performed based on the annotation manual (see Appendix), using CAT. Before the test, a pilot study was carried out with the twofold purpose of (i) calibrating the annotation guidelines, and (ii) training on the guidelines and the use of CAT. During the pilot test, the annotators independently tagged a small sample of texts with the same characteristics of those included in the corpus. This was followed

by a discussion session, during which the coders compared their annotations and reconciled disagreements, to the extent possible. Following Wiebe et al. (2005), the reconciled annotations from the pilot test were used as a gold standard reference for the subsequent annotations. After the training phase, the test took place. The coders annotated random paragraphs independently, amounting to approximately 25% of the corpus. The paragraphs were randomly selected using an automatic procedure.[6] Similarly to Read and Carroll (2010), the annotation was performed over three rounds, punctuated by an intermediate phase of analysis and reconciliation of disagreements. These intermediate sessions also served to address unanticipated annotation problems and further refine the annotation guidelines. Inter-coder agreement, both prior to and after reconciliation, was measured and recorded. In total, the test required approximately sixty person-hours to complete.[7] The next section reports the results of the test.

5.3.2 Inter-coder agreement results

For both the markable identification and classification tasks, we report the agreement scores obtained prior as well as after reconciliation. However, it is important to note that only the scores obtained before reconciliation count as reliability data proper (Krippendorff, 2004: 219). Nevertheless, the agreement scores achieved through discussion and reconciliation provide a useful indication of how much agreement can be reached when collaboratively annotating our corpus, and when inaccuracies and inconsistencies due to annotator fatigue and distraction are filtered out from the data.

Table 2. Inter-coder agreement results: markable identification task

The results of the markable identification task are reported in table 2. Following Read and Carroll (2010) and Taboada and Carretero (2012), to measure the amount of agreement between annotators on this type of task we use *precision* (PRE), *recall* (REC) and *F measure* (F-score) scores. *Kappa* statistics (Cohen et al., 1960) are not suitable for this task, as it focuses on the identification of markables rather than the labeling of units of fixed length. If we take the annotations produced by author 1 as

the gold standard (i.e. the 'correct' annotations):

- PRE indicates the fraction of units identified by author 2 that are actually *relevant* (or 'correct'), i.e. that have also been annotated by author 1. It is calculated dividing the number of units identified by both annotators by the number of units annotated by author 2.

- REC represents the fraction of relevant units that have been successfully identified by author 2. It is calculated dividing the number of relevant units identified by annotator 2 (i.e. those that have also been annotated by author 1) by the total number of relevant units, i.e. the number of units annotated by author 1 (i.e. the gold standard).

- F-score provides a synthetic measure of PRE and REC. It is calculated as the weighted harmonic mean of precision and recall.

For the sake of simplicity and following Read and Carroll (2010) and Wiebe et al. (2005), we consider overlapping annotated text units as matches.[8]

As table 2 shows, inter-coder agreement for the markable identification task before reconciliation is substantial, with an overall mean F-score of 0.79. The scores obtained from this test are higher than those reported in Read and Carroll (2010) and similar to those obtained by Taboada and Carretero (2012). The table also shows that, while agreement for the categories of AFFECT and ENGAGEMENT is robust, JUDGEMENT was a more problematic category. This is not surprising. As mentioned earlier, in fact, one of the primary communicative goals of the texts analyzed is precisely that of conveying a positive evaluation of the company. In a highly evaluative and positive context, discerning explicit from invoked positive JUDGEMENT and identifying bounded evaluative units turned out to be a very complex task. Notwithstanding, the scores obtained are higher than those reported for this category in Read and Carroll (2010).

Table 2 also shows the level of agreement reached after thoroughly discussing discrepancies between the annotations. After reconciliation, agreement becomes almost perfect, with an F-score of 0.98. The discussion and reconciliation sessions allowed not only to collaboratively identify the best solution for the disagreements, but it also highlighted various instances of inconsistencies and of unproblematic annotations that were missed due to fatigue or distraction. These issues, which are an

inevitable drawback of manual annotation, have clearly affected the inter-coder agreement scores obtained before reconciliation.

Table 3 reports the results for the classification task. The table shows the levels of agreement reached in the classification of the three main APPRAISAL categories included in the coding scheme, at two levels of granularity. The scores in the table refer to the *observed agreement* between the annotators, both before and after reconciliation. Observed agreement is obtained by dividing the total number of judgments on which the annotators agree by the total number of markable items, i.e. the number of instances that were identified by both annotators independently. Observed agreement is not a very robust measure of agreement, as it does not account for chance. While Cohen's kappa would be preferable, the highly skewed distribution of the categories in our data, where a disproportionate number of instances fall under one category, leads to the well-known 'paradox' that high levels of observed agreement correspond to very low kappa scores (Artstein and Poesio, 2008; Di Eugenio and Glass, 2004). To avoid this problem, we report only observed agreement scores, with the warning that these figures are not corrected for chance and are, therefore, relatively poor indicators of reliability.

Table 3. Inter-coder agreement results: classification task

Table 3 shows very high levels of observed agreement on the classification task at the lower level of granularity, with all the coefficients above 0.90. Predictably, the scores decrease when the analysis becomes more fine-grained and more options are available (note that only one level of granularity was considered for AFFECT; cf. the coding scheme in Fig. 3). Agreement for the category of JUDGEMENT at this level of granularity is particularly low, which is due to the ambiguity and under-specification of the classification criteria for these subcategories in the APPRAISAL literature, and to the fact that many instances of JUDGEMENT were found to be compatible with more than one reading/label. As shown in the table, upon reconciliation most disagreements could be solved and observed agreement neared 100%.

In sum, the inter-coder agreement test we carried out on a sample from the corpus indicates substantial agreement between independent annotators in the tasks of

identifying and classifying instances of APPRAISAL in our corpus, based on the guidelines designed for this study (see Appendix).

6. Analysis

As discussed above, Fuoli and Paradis (2014) propose that, when a trust-breaking event occurs, the trust-breaker's ability, integrity and benevolence may be at stake. Accordingly, the trust-breaker's discourse will be geared towards reshaping the trust-giver's impressions of their trustworthiness along these three dimensions, with the ultimate goal of repairing trust.

Based on the events that followed the Deepwater Horizon oil spill and the public reactions to BP's handling of the crisis, we may infer that, after the accident, the company's trustworthiness was at stake on all levels. BP's ability was being questioned due to its repeated failed attempts to control the spill.[9] Its integrity was undermined by its initial attempts to downplay its effects (Webb, 2010) and its biased reports of the amount of oil gushing into the sea (O'Connor, 2011; Shogren, 2011). BP's benevolence came under intense scrutiny after some controversial public statements by the company's CEO ('I want my life back') and President ('We care about the small people'), which cast doubts on BP's real priorities and created the impression that it did not truly care about the people affected by the accident (O'Connor, 2011).

In this section, we combine quantitative and qualitative analysis of the use of APPRAISAL resources in BP's CEO letters to investigate how ability, integrity and benevolence are discursively (re-)constructed. The analysis addresses the following questions:

1. Does the frequency of AFFECT, JUDGEMENT and ENGAGEMENT expressions change in BP's discourse after the spill? If so, how?

2. How are these resources deployed in BP's letters to repair trust in the company's ability, integrity and benevolence? What aspects of the company's trustworthiness are emphasized?

3. Is there any difference between the trust-repair strategies adopted in the 2010 and 2011 letters?

In section 6.1, the distribution of the APPRAISAL categories considered is explored in order to identify patterns of change in BP's discourse after the spill. The patterns identified are used as a guide for the qualitative analysis of BP's texts, which is presented in section 6.2. In section 6.3, the trust-repair strategies that emerge from the analysis are discussed.

6.1 Quantitative analysis

In order to identify significant changes in the use of the APPRAISAL resources considered in BP's post-crisis discourse, the BP letters published after the spill were compared with those published before it and by the other oil companies during the same time span. As the other companies were only indirectly affected by the spill, their texts were treated as a baseline.

The frequency of expressions of AFFECT, ENGAGEMENT and JUDGEMENT in each text included in the corpus was compared using log-linear analysis. The frequency data were classified in a three-way contingency table according to three factors: APPRAISAL category, company and year. A log-linear model including the main effects of these three factors was fitted to the data to calculate the expected frequency counts. These values were compared to the observed frequencies.[10]

The mosaic plot in Fig. 4 shows the results of the analysis. It schematically represents the observed frequency of the annotated features (cell size) and marks the statistically significant deviations from the expected frequencies through a shade/outline scheme. Shaded cells with solid outline denote significantly overrepresented values, while shaded cells with dashed outline correspond to significantly underrepresented values. The raw frequencies of the different appraisal types, which were used for the log-linear analysis, are reported in the boxes.[11]

Figure 4. Mosaic plot: Analysis results. Standardized residuals (z-scores) are a statistical measure of the difference between observed and expected frequencies. Values higher than +2 or smaller than -2 indicate that this difference is significant.

From the results presented in Fig. 4 it becomes apparent that the use of APPRAISAL resources in BP's CEO letters changed after the spill. The plot shows that AFFECT and ENGAGEMENT are both significantly overrepresented in BP's letters after the accident. The frequency of ENGAGEMENT markers in the 2010 letter shows a greater deviation from the expected values, with standardized residuals larger than 4. Conversely, JUDGEMENT is significantly underrepresented in the letter released the year following the accident. The letters published by BP after the spill thus appear to place greater emphasis on emotions and to shift towards a more overtly dialogic style. The pattern observed for JUDGEMENT suggests that, in the letter published the year after the spill, the CEO adopts a more neutral, factual style, with comparatively fewer explicit positive assessments of the company's performance and qualities. Notably, similar patterns cannot be observed for any of the other companies considered.

The results of the quantitative analysis provide useful insights into the CEO's discursive response to the spill, highlighting patterns of change and salient phenomena in the company's discourse after the accident. These results can be used as a guide for a more fine-grained, qualitative analysis of BP's texts, which should provide a more complete picture of the function of the annotated features in context, and of the discourse strategies adopted by the CEO in the attempt to restore trust in BP. It is to this type of analysis that we now turn.

6.2 Qualitative analysis

As shown in Fig. 4, AFFECT is overrepresented in both BP's 2010 and 2011 letter. Unsurprisingly, most of the instances of this category found in the 2010 letter are used to convey negative emotions. They are placed in a very prominent position, at the beginning of the text, as shown in (14).

(14) Dear fellow shareholder,

The tragic[12] events of 2010 will forever be written in the memory of this company and the people who work here. The explosion and fire on the Deepwater Horizon rig shocked everyone within BP, and we feel great sadness[13] that 11 people died. We are deeply sorry for the grief felt by their families and friends. We know nothing can restore the loss of those men.

The expressions of AFFECT included in this excerpt (underlined items) refer to both

the company's emotional reaction to the accident as well as to the emotions felt by the families and friends of the victims. These expressions serve the twofold purpose of conveying BP's affective stance towards the accident, and to acknowledge the suffering caused by it. In this sense, the ultimate communicative function of these expressions in this particular context can be seen as that of showing empathy and care for the people affected by the spill. These feelings are further stressed in the following paragraph, where the CEO foregrounds his personal attachment to the areas hit by the spill by referring to his roots in one of the affected states and recalling fond childhood memories on the Gulf coast.

(15) And it all started in a part of the world that's very close to my heart. I grew up in Mississippi, and spent summers with my family swimming and fishing in the Gulf. I know those beaches and waters well. When I heard about the accident I could immediately picture how it might affect the people who live and work along that coast.

If we interpret this passage in the light of the widespread resentment caused by some controversial declarations by the company's former *foreign* CEO and President in the aftermath of the spill (see section 2), it is possible to see the CEO's emphasis on his American identity and affective attachment to the Gulf (cf. *very close to my heart*) as a strategic move aimed at re-negotiating solidarity and alignment with the victims of the spill and, at the same time, counteracting the widely held impression that the company did not truly care about them. In other words, the CEO's personal bond to the Gulf Coast can be taken as a 'guarantee' that the company of which he is the leader truly cares about the people affected by the spill, and will strive to protect them and the Gulf Coast's environment from its negative effects.

One conclusion that emerges from the discussion above is that, in the 2010 letter, AFFECT plays a key role in the CEO's attempt to communicate empathy, solidarity and care towards the victims of the spill. From a quantitative point of view, AFFECT is a prominent category also in the 2011 letter. In this text, however, expressions of AFFECT perform a different communicative function. Notably, while the vast majority of the instances of AFFECT found in the 2010 letter have a negative valence, the 2011 text includes a comparatively higher number of instances of positive AFFECT. AFFECT is used here to foster confidence in the company's recovery and future prospects. The CEO declares, for instance, to be *proud* of how BP responded to the accident, *pleased* about several legal settlements reached during the preceding year, and *confident* about BP's 'ability to design, engineer and operate large installations safely'.

As shown in Fig. 4, ENGAGEMENT is significantly over-represented in both BP's letters after the spill and is particularly prominent in the 2010 letter. This finding indicates that the letters published after the accident are more overtly dialogic, i.e. the CEO explicitly engages with alternative propositions, opinions and voices more often. This tendency is particularly evident in the 2010 letter, where the CEO responds to several stakeholders' concerns and criticisms. Towards the end of the text, for instance, he entertains an imaginary dialogue with the stakeholders and seeks to reassure them that the company 'has learned its lesson from the spill'.

(16) I have heard people ask 'Does BP 'get it' ?" Residents of the Gulf, our employees and investors, governments, industry partners and people around the world all want to know whether we understand that a return to business-as-usual is not an option. We may not have communicated it enough at times, but yes, we get it. Our fundamental purpose is to create value for shareholders, but we also see ourselves as part of society, not apart from it.

A striking observation that can be made from example (16) is that all the markers of ENGAGEMENT identified (underlined items) are dialogically contracting, i.e. they act to reject or disallow alternative viewpoints and opinions. The leading question and answer pair *Does BP 'get it' ? Yes, we get it*[14], the three instances of negation/denial and the counter-expectation marker *but* all contribute to 'fending off' and pre-empting adverse opinions and discourses. This pattern characterizes the whole 2010 letter, where dialogically contracting ENGAGEMENT markers are repeatedly used to reject or pre-empt potential criticisms of the company's decisions and actions, and protect the company's integrity from further damage.

ENGAGEMENT markers in the 2011 letter are also, for the vast majority, dialogically contracting. However, in this letter they perform a different communicative function. ENGAGEMENT resources are used by the CEO to boost the credibility of his promises and predictions about the future of the company, rather than to respond to criticisms and concerns. A clear example of this is reported in (17).

(17)

- First and foremost, you will see a continuing, relentless focus on safety and risk management.

- You will see the company play to its strengths […].

- You will see a company that is simpler and more focused […].
- You will see a company that is organized effectively and applies its standards consistently.
- You will see more visibility from us on our individual businesses.
- You will be able to measure the effects of active portfolio management […].
- You will be able to measure the contribution of new upstream projects with higher margins, as they come onstream over the next three years.

The repeated, emphatic use of the evidential verbs *see* and *measure* in (17) has a strong dialogically contracting effect, as these verbs act to rule out any doubt or concern that any scenario other than that envisaged by the CEO will occur. Put differently, the reader should trust the CEO's statements, as they will be able to check them against tangible evidence. The reassuring, trust-building force of these utterances is bolstered by the repeated use of the second person pronoun *you*, which construes a direct, 'face-to-face' relationship between the CEO and the addressees and indexes honesty and personal commitment. Interestingly, the second person pronoun is never found in the other BP letters and is extremely rare in the other companies' texts. In sum, ENGAGEMENT resources are more frequently used in BP's letters after the accident. In the 2010 letter they mainly serve to protect the company's integrity, whereas in the 2011 text they are primarily used to boost shareholders' confidence in BP's recovery.

As the quantitative analysis has shown, JUDGEMENT is significantly under-represented in BP's 2010 letter. This result suggests that the evaluative force of this text is comparatively down-toned. Explicit positive assessments of the company's behavior and performance are rare, and positive evaluations tend to be invoked, rather than inscribed. Where positive evaluative expressions are used, they serve to stress BP's resilience to the crisis, as for example in (18).

(18) The sound underlying performance across our business continues to give us a solid foundation, and speaks volumes for the inner strengths of BP and our people.

The relative backgrounding of positive evaluation in BP's 2010 letter becomes strikingly evident if we compare this text with the letter published by one of the

company's main competitors, ExxonMobil, in the same year.

(19) To Our Shareholders

> ExxonMobil continues to deliver superior long term shareholder value. We succeed by upholding the values that set us apart: a commitment to safety, operational excellence, and risk management; a disciplined, long term approach to investing; and the development and application of advanced technology and innovation.

As Fig. 4 shows and example (19) confirms, positive JUDGEMENT in ExxonMobil's 2010 letter is strongly foregrounded. ExxonMobil's CEO adopts a highly evaluative writing style, characterized by a repeated use of positive evaluative expressions that emphasize the company's qualities and competitiveness. Interestingly, ExxonMobil was responsible in 1989 for one of the most catastrophic oil spills in history, Alaska's Exxon Valdez oil spill. In relation to this, it is noteworthy to observe that the quantitative pattern for the use of APPRAISAL resources in ExxonMobil's 2010 letter is opposite to that observed for BP. ENGAGEMENT is significantly underrepresented in ExxonMobil's text, configuring a very assertive, monologic writing style, whereas JUDGEMENT is overrepresented. Considering that ExxonMobil was frequently mentioned in the news in the aftermath of the Deepwater Horizon accident, when comparisons were often drawn between the two spills, we might read these patterns as indicating ExxonMobil's strategic attempt to discursively differentiate itself from the competitor, in order to protect its image from the collateral damage that the BP spill may cause.

6.3 BP's trust-repair strategies

The analysis presented above has highlighted clear patterns of change in the use of APPRAISAL resources in BP's CEO letters following the Deepwater Horizon oil spill. In this section, we interpret the results of the analysis in light of Fuoli and Paradis' (2014) model and seek to provide a concise answer to our research questions.

As shown above, BP's 2010 letter places substantial emphasis on emotions. Expressions of AFFECT are primarily used to show empathy and care towards the victims of the spill. In terms of Fuoli and Paradis' (2014) model, we may thus conclude that the use of AFFECT resources in this text is geared towards constructing

benevolence. The highly dialogic nature of the 2010 text, evidenced by the higher-than-expected use of ENGAGEMENT expressions, can be explained in light of the numerous controversies and concerns generated by the accident. The CEO expends considerable effort in trying to neutralize the negative discourses about the company that circulated in the aftermath of the accident, and ENGAGEMENT resources play a central role in this endeavor. The use of ENGAGEMENT resources in the 2010 letter may thus be seen as configuring a defensive discourse strategy aimed at protecting the company's integrity. Finally, the underuse of explicit positive evaluative language may also be seen as contributing to constructing integrity. Indeed, given the magnitude of the effects of the spill, frequent explicit positive self-assessments may have come across as inappropriate and insincere, further undermining BP's trustworthiness. The adoption of a relatively more factual and objective tone may thus be read as an attempt to communicate humbleness and credibility. The backgrounding of explicit positive evaluation may also be seen to imply that the ability facet of BP's trustworthiness is de-emphasized. Therefore, we may conclude that, in the 2010 letter, BP's CEO adopts a trust-repair strategy that emphasizes the company's benevolence and integrity, while down-toning ability.

The picture that emerges from the analysis of the 2011 letter is rather different. Ability is more strongly emphasized in this text, and optimism in the company's recovery and future performance is promoted by the CEO. As far as the integrity facet of BP's trustworthiness is concerned, credibility appears to be the main aspect in focus. The CEO aims to reassure the shareholders that the company will be able to overcome the crisis and return to profitability. Through the use of dialogically contracting expressions of ENGAGEMENT, he seeks to boost the credibility of his predictions, and remove uncertainties about the future of the company. Compared to the 2010 letter, the benevolence facet of BP's trustworthiness appears to be less explicitly foregrounded in this text. Therefore, the 2011 letter shows a change in the CEO's trust-repair discourse strategy, with a stronger focus on the company's ability and a significant communicative effort to restore and promote shareholders' confidence.

7. Conclusion

Manual corpus annotation allows for exhaustive and detailed corpus-based analyses of evaluative language that would not be possible with purely automatic techniques,

given the complex and context-dependent nature of evaluation in discourse. As discussed above, transparency, reliability and replicability are crucial issues that need to be addressed when taking this kind of approach. We propose that these aspects can be optimized by adopting explicit annotation guidelines and by assessing and reporting inter-coder agreement. The results of the intercoder agreement test reported above are encouraging, as they show that, in spite of the complexity of the task at hand, robust agreement between independent coders can be reached, given explicit guidelines and appropriate training.

Clearly, the approach discussed here has several limitations, the most important being scalability, i.e. the extent to which it can be applied to increasingly larger corpora. Compared to purely automatic corpus methods, this type of approach is relatively time consuming and resource-intensive. As a consequence, the amount of data that can be processed is, in normal circumstances, more limited than with traditional automatic techniques. We suggest that scalability may be improved by (i) using random sampling techniques as opposed to annotating full texts, and (ii) simplifying the coding scheme so as to make the manual annotation process faster. Another approach to address these limitations is to develop new, advanced software tools to assist, simplify and quicken the process of manual corpus annotation. We hope that this article will inspire more research and development in this area.

Acknowledgments

We would like to thank Joost van de Weijer (Lund University, Sweden) for his support with statistics and helpful comments on an earlier version of this article. We are also grateful to Carita Paradis (Lund University, Sweden) for her invaluable guidance throughout the preparation of the article. Thanks also go to Stina Ericsson (Linnaeus University, Sweden) and two anonymous reviewers for their constructive comments on the manuscript.

Notes

[1] BP's 2008 and 2009 annual reports did not include a letter to shareholders signed by the CEO, but instead featured an interview with him. Given that the interview genre is radically different from that of CEO letters, we included in the corpus the letters to shareholders signed by the company's Chairman.

[2] Clearly, one could repeatedly shift between word or n-gram lists and the original texts to study the context-specific meaning of evaluative expressions, but that seems unpractical, and defeats the purpose of using automatic techniques.

[3] In the analysis presented here, all evaluative expressions belonging to selected APPRAISAL categories are identified and classified. The analysis is therefore limited in scope, but still exhaustive in the sense that it covers *all* expressions instantiating the categories considered, and is not limited to a pre-determined set of forms.

[4] CAT can be accessed, free of charge, at this url: https://dh.fbk.eu/resources/catcontent-annotation-tool. An alternative and widely used annotation program is the UAM Corpus Tool (O'Donnell, 2008). Some scholars (e.g. Bednarek, 2008) have used the Altova XMLSpy editor (http://www.altova.com/xmlspy.html). This tool, however, is not freely available and is not specifically designed for annotating corpora, which makes it less intuitive and easy to use compared to UAM or CAT. For a review of different manual corpus annotation tools, see O'Donnell (2014).

[5] The annotation scheme diagram was generated using the UAM Corpus Tool (O'Donnell, 2008).

[6] The random paragraphs were selected using the a freely available online tool called *Random Line Picker*. The tool is available at http://textmechanic.com/Random-Line-Picker. html

[7] This figure accounts for the time needed to complete the annotation tasks, discussing and reconciling the disagreements and analyzing the agreement data. It sums the time spent by both annotators, thus on average each annotator devoted approximately 30 hours of their time to the test. Author 1, however, was in charge of data preparation and analysis, so he devoted proportionally more time to the test.

[8] Initial efforts were spent to try to reach agreement on boundary placement, but this turned out to be an impracticable task.

[9] For a detailed account of all the attempts at stopping the spill, see http://www.nytimes.com/interactive/2010/05/25/us/20100525-topkill-diagram.html?_r=0.

[10] All statistical analysis was performed using R, version 3.0.1 (http://www.R-project.org/).

[11] Differences in text length are not problematic for the analysis.

[12] The word *tragic* in (14) was annotated as an instance of COVERT AFFECT (Bednarek, 2009).

[13] Even though GRADUATION was not included in the analysis, GRADUATION markers such as *deeply* or *great* were by convention annotated as belonging to an AFFECT or JUDGEMENT unit, when used as modifiers. This choice has no impact on the quantitative figures. This choice is accounted for in the annotation manual.

[14] The question-answer pair was annotated as a single unit of ENGAGEMENT:CONTRACT:DISCLAIM.

Appendix

The annotation manual can be found as supplementary material to the web-based version of this article.

References

Artstein, R. and Poesio, M. 2008. 'Inter-coder agreement for computational linguistics', *Computational Linguistics* 34(4), pp 555–596.

Bakhtin, M. 1981. *The dialogical imagination*. Austin: University of Texas Press.

Barney, J. B. and Hansen, M. H. 1994. 'Trustworthiness as a source of competitive advantage', *Strategic management journal* 15(1), pp 175–190.

Bartalesi Lenzi, V., Moretti, G., and Sprugnoli, R. 2012. 'CAT: the CELCT Annotation Tool', in *LREC 2012 proceedings*, pp 333–338.

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

Bednarek, M. 2008. *Emotion talk across corpora*. Houndmills, Basingstoke: Palgrave Macmillan.

Bednarek, M. 2009. 'Language patterns and attitude', *Functions of language* 16(2), pp 165–192.

Benoit, W. L. 1997. 'Image repair discourse and crisis communication', *Public relations review* 23(2), pp 177–186.

Bhatia, V. 2004. *Worlds of written discourse: A genre-based view*. New York: Continuum.

Biber, D. 2006. 'Stance in spoken and written university registers', *Journal of English for Academic Purposes* 5(2), pp 97–116.

Biber, D. and Finegan, E. 1988. 'Adverbial stance types in English', *Discourse processes* 11(1), pp 1–34.

Biber, D. and Finegan, E. 1989. 'Styles of stance in English: Lexical and grammatical marking of evidentiality and affect', *Text-Interdisciplinary Journal for the Study of Discourse* 9(1), pp 93–124.

Breeze, R. 2012. 'Legitimation in corporate discourse: Oil corporations after deepwater horizon', *Discourse & Society* 23(1), pp 3–18.

Breeze, R. 2013. *Corporate Discourse*. London & New York: Bloomsbury Academic.

Camiciottoli, B. C. 2013. *Rhetoric in financial discourse: A linguistic analysis of ICT-mediated disclosure genres*. Amsterdam & New York: Rodopi.

Carretero, M. and Taboada, M. 2014. 'Graduation within the scope of Attitude in English and Spanish consumer reviews of books and movies', in Thompson, G. and Alba-Juez, L. (eds.), *Evaluation in context*, pp 221–239. Amsterdam and Philadelphia: John Benjamins.

Cohen, J. 1960. 'A coefficient of agreement for nominal scales', *Educational and psychological measurement* 20(1), pp 37–46.

Conrad, S. and Biber, D. 2000. 'Adverbial marking of stance in speech and writing', in Hunston, S. and Thompson, G. (eds.) *Evaluation in text: Authorial stance and the construction of discourse*, pp 56–73. Oxford: Oxford University Press.

Cook, K. S. 2001. *Trust in society*. New York: Russell Sage Foundation.

Di Eugenio, B. and Glass, M. 2004. 'The kappa statistic: A second look', *Computational linguistics* 30(1), pp 95–101.

Eggins, S. 2004. *An introduction to systemic functional linguistics*. Continuum International Publishing Group.

Fuoli, M. 2012. 'Assessing social responsibility: A quantitative analysis of APPRAISAL in BP's and IKEA's social reports', *Discourse & Communication* 6(1), pp 55–81.

Fuoli, M., & Paradis, C. (2014). 'A model of trust-repair discourse', *Journal of Pragmatics* 74, pp 52-69.

García-Marza, D. 2005. 'Trust and dialogue: theoretical approaches to ethics auditing', *Journal of Business Ethics* 57(3), pp 209–219.

Garzone, G. 2005. 'Letters to shareholders and chairman's statements: textual variability and generic integrity', in Gillaerts, P. and Gotti, M. (eds.) *Genre variation in business letters*, pp 179–204. Bern: Peter Lang.

Gillaerts, P. and Van de Velde, F. 2011. 'Metadiscourse on the move: the CEO's letter revisited', in Grazone, G. and Gotti, M. (eds.) *Discourse, communication and the enterprise: genres and trends*, pp 151–168. Bern: Peter Lang.

Halliday, M. 1994. *An Introduction to Functional Grammar*. London: Edward Arnold.

Harlow, W. F., Brantley, B. C., and Harlow, R. M. 2011. 'BP initial image repair strategies after the *Deepwater Horizon* spill', *Public Relations Review* 37(1), pp 80–83.

Hommerberg, C. and Don, L. Forthcoming. 'APPRAISAL and the language of wine APPRECIATION'. Functions of language.

Hunston, S. 2004. 'Counting the uncountable: Problems of identifying evaluation in a text and in a corpus', in Partington, A., Morley, J., and Haarman, L. (eds.) *Corpora and discourse*, pp 157-188. Bern: Peter Lang.

Hunston, S. 2011. *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York and London: Routledge.

Hunston, S., and Sinclair, J. (2000). 'A local grammar of evaluation', in Hunston, S. and Thompson, G. (eds.) *Evaluation in text: Authorial stance and the construction of discourse*, pp 74–101. Oxford: Oxford University Press.

Hunston, S. and Thompson, G. 2000. *Evaluation in text: Authorial stance and the construction of discourse*. Oxford: Oxford University Press.

Hyland, K. 1998. 'Exploring corporate rhetoric: metadiscourse in the CEO's letter', *Journal of Business Communication* 35(2), pp 224–244.

Hyland, K. 2005. *Metadiscourse: Exploring interaction in writing*. London & New York: Continuum.

Ingenhoff, D. and Sommer, K. 2010. 'Trust in companies and in CEOs: A comparative study of the main influences', *Journal of business ethics* 95(3), pp 339–355.

Kaltenbacher, M. 2006. 'Culture related linguistic differences in tourist websites: the

emotive and the factual–A corpus analysis within the framework of APPRAISAL', in Thompson, G. & Hunston, S. (eds.) *System and Corpus–Exploring Connections*, pp 269–292. London: Equinox.

Krippendorff, K. 2004. *Content analysis: An introduction to its methodology*. Thousand Oaks, London and New Deli: Sage.

Linell, P. and Markova, I. 2013. *Dialogical Approaches to Trust in Communication*. Information Age Publishing.

Lipovsky, C. 2008. 'Constructing affiliation and solidarity in job interviews', *Discourse & Communication* 2(4), pp 411–432.

Lipovsky, C. (2013). 'Negotiating ones expertise through appraisal in CVs'. *Linguistics and the Human Sciences* 8(3), pp 307–333.

Mackay, J. and Parkinson, J. 2009. ' "My very own mission impossible: an APPRAISAL analysis of student teacher reflections on a design and technology project', *Text & Talk* 29(6), pp 729–753.

Markova, I. and Gillespie, A. 2008. *Trust and Distrust: Sociocultural Perspectives*. Information Age Publishing.

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

Mayer, R., Davis, J., and Schoorman, F. 1995. 'An integrative model of organizational trust', *Academy of management review* 20(3), pp 709–734.

Mackay, J. and Parkinson, J. 2009. ' "My very own mission impossible": an APPRAISAL analysis of student teacher reflections on a design and technology project', *Text & Talk* 29(6), pp 729–753.

Muralidharan, S., Dillistone, K., and Shin, J.-H. 2011. 'The gulf coast oil spill: Extending the theory of image restoration discourse to the realm of social media and beyond petroleum', *Public Relations Review* 37(3), pp 226–232.

Murphy, A. C. (2013). 'On "true" portraits of Letters to Shareholders–and the importance of phraseological analysis', *International Journal of Corpus Linguistics* 18(1), pp 57-82.

O'Connor, E. O. 2011. 'Organizational Apologies: BP as a Case Study', *Vanderbilt Law Review* 64, pp 1959–1991.

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

O'Donnell, M. 2014. 'Exploring identity through appraisal analysis: A corpus annotation methodology', *Linguistics and the Human Sciences* 9(1), pp 95-116.

Pirson, M. and Malhotra, D. 2011. 'Foundations of organizational trust: What matters to different stakeholders?', *Organization Science* 22(4), pp 1087–1104.

Poppo, L. and Schepker, D. J. 2010. 'Repairing public trust in organizations', *Corporate Reputation Review* 13(2), pp 124–141.

Pounds, G. 2010. 'Attitude and subjectivity in Italian and British hard-news reporting: The construction of a culture-specific reporter voice', *Discourse Studies* 12(1), pp 106–137.

Pounds, G. 2011. ' "This property offers much character and charm": evaluation in the discourse of online property advertising', *Text & Talk* 31(2), pp 195–220.

Read, J. and Carroll, J. 2010. 'Annotating expressions of appraisal in English', *Language Resources and Evaluation* 46, pp 421–447.

Ryshina-Pankova, M. (2014). 'Exploring argumentation in course-related blogs through ENGAGEMENT', in Thompson, G. and Alba-Juez, L. (eds.), *Evaluation in context*, 281–302. Amsterdam and Philadelphia: John Benjamins.

Robertson, C. and Krauss, C. 2010. 'Gulf spill is the largest of its kind, scientists say', *The New York Times online*. Retrieved from http://www.nytimes.com/. Accessed 14 February 2014.

Santamaría-García, C. (2014). 'Evaluative discourse and politeness in university

students' communication through social networking sites', in Thompson, G. and Alba-Juez, L. (eds.), *Evaluation in context*, 387–411. Amsterdam and Philadelphia: John Benjamins.

Schultz, F., Kleinnijenhuis, J., Oegema, D., Utz, S., and van Atteveldt, W. 2012. 'Strategic framing in the BP crisis: A semantic network analysis of associative frames', *Public Relations Review* 38(1), pp 97–107.

Shogren, E. 2011. 'BP: A Textbook Example Of How Not To Handle PR', Retrieved from http://www.npr.org.

Spooren, W. and Degand, L. 2010. 'Coding coherence relations: Reliability and validity', *Corpus Linguistics and Linguistic Theory* 6(2), pp 241–266.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. 2011. 'Lexicon-based methods for sentiment analysis', Computational linguistics 37(2), pp 267–307.

Taboada, M. and Carretero, M. 2012. 'Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish', *Linguistics & the Human Sciences* 6, pp 275–295.

Thompson, G. and Alba-Juez, L. 2014. *Evaluation in context*. Amsterdam and Philadelphia: John Benjamins.

Voormann, H. and Gut, U. 2008. 'Agile corpus creation', *Corpus Linguistics and Linguistic Theory* 4(2), pp 235–251.

Webb, T. 2010. 'BP boss admits job on the line over Gulf oil spill', *The Guardian online*. Retrieved from http://www.theguardian.com/business/. Accessed 14 February 2014.

White, P. 2003. 'Beyond modality and hedging: A dialogic view of the language of intersubjective stance', *Text-Interdisciplinary Journal for the Study of Discourse* 23(2), pp 259–284.

White, P. R. 2012. 'Exploring the axiological workings of 'reporter voice' news stories–attribution and attitudinal positioning', *Discourse, Context & Media*

1(2-3), pp 57–67.

Wickman, C. 2013. 'Rhetorical framing in corporate press releases: The case of British petroleum and the gulf oil spill', *Environmental Communication* 8(1), pp 3–20.

Wiebe, J., Wilson, T., and Cardie, C. 2005. 'Annotating expressions of opinions and emotions in language', *Language Resources and Evaluation* 39(2), pp 165–210.

Figures and tables

Table 1. Corpus details (number of words per text)

		Companies				
		BP (UK)	*Chevron (US)*	*ConocoPhillips (US)*	*ExxonMobil (US)*	*Royal Dutch Shell (UK)*
Year	*2008*	1113	839	1910	964	886
	2009	1085	897	850	876	1014
	2010	2118	993	1325	935	1326
	2011	1840	942	1758	877	1300

Figure 1. The CAT user interface

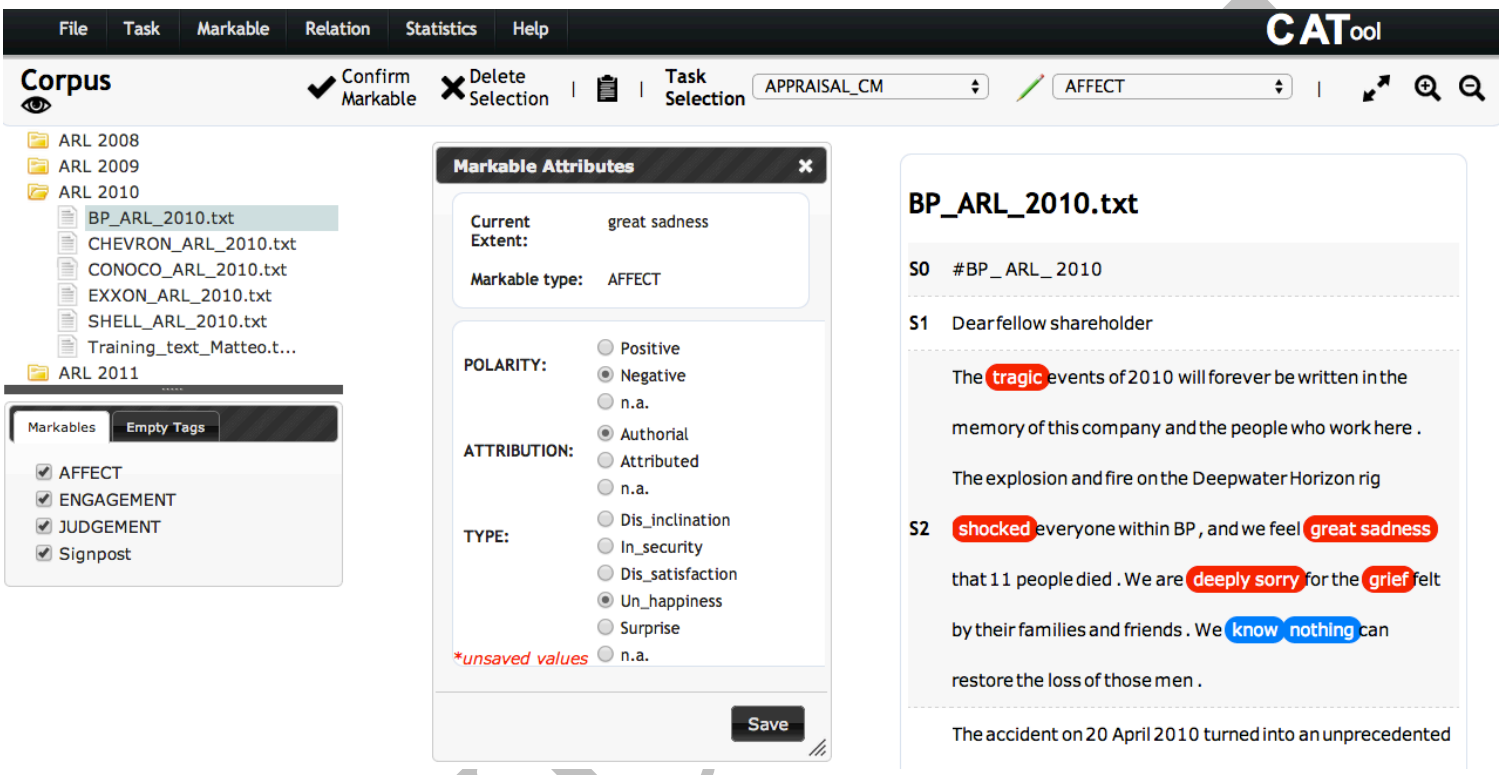

Figure 2. The CAT output

Tokens	TYPE	POLARITY	ATTRIBUTION	Comments
proud	Dis_satisfaction	Positive	Authorial	
want	Dis_inclination	Positive	Attributed	
deeply appreciate	Dis_satisfaction	Positive	Authorial	
inspire	Dis_inclination	Positive	Authorial	
admired	Dis_satisfaction	Positive	Attributed	
confident	In_security	Positive	Authorial	
tragic	Un_happiness	Negative	n.a.	covert affect
shocked	Surprise	Negative	Authorial	negative surprise?
great sadness	Un_happiness	Negative	Authorial	
deeply sorry	Un_happiness	Negative	Authorial	
grief	Un_happiness	Negative	Attributed	

Figure 3. Annotation scheme

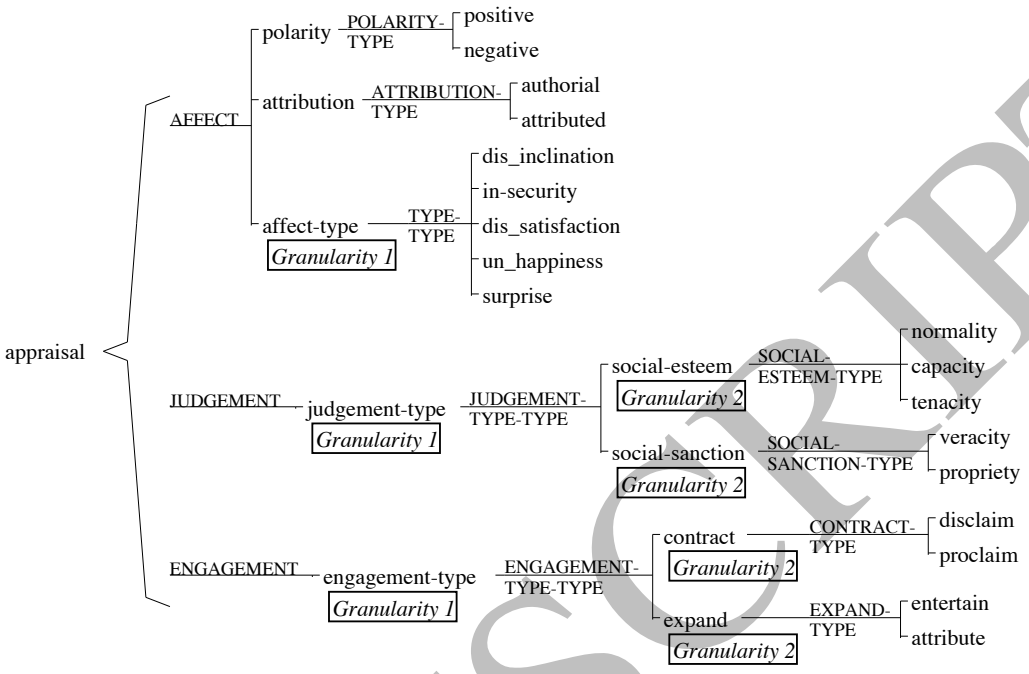

Table 2. Inter-coder agreement results: markable identification task

	PRE		REC		F-score	
	Before	After	Before	After	Before	After
AFFECT	0.82	1.00	0.90	1.00	0.86	1.00
JUDGEMENT	0.71	0.92	0.71	0.98	0.71	0.95
ENGAGEMENT	0.77	0.97	0.85	1.00	0.81	0.98
Mean	0.77	0.96	0.82	0.99	**0.79**	**0.98**

Table 3. Inter-coder agreement results: classification task

	Observed agreement			
	Granularity 1		Granularity 2	
	Before	After	Before	After
AFFECT	0.93	0.96	n.a.	n.a.
JUDGEMENT	0.94	0.98	0.75	0.92
ENGAGEMENT	0.91	1.00	0.88	1.00
Mean	0.93	0.98	0.81	0.96

Figure 4. Analysis results