A step-wise method for annotating APPRAISAL

Fuoli, Matteo

Published in:
Functions of Language

DOI:
10.1075/fol.15016.fuo

2018

Document Version:
Peer reviewed version (aka post-print)

Link to publication

Citation for published version (APA):
Fuoli, M. (2018). A step-wise method for annotating APPRAISAL. *Functions of Language*, *25*(2), 229 - 258 . https://doi.org/10.1075/fol.15016.fuo

Total number of authors:
1

A step-wise method for annotating APPRAISAL

Matteo Fuoli
Centre for Languages and Literature
Lund University

Abstract

Despite a growing awareness of methodological issues, the literature on APPRAISAL has not so far provided adequate answers to some of the key challenges involved in reliably identifying and classifying evaluative language expressions. This article presents a step-wise method for the manual annotation of APPRAISAL in text that is designed to optimize reliability, replicability, and transparency. The procedure consists of seven steps, from the creation of a context-specific annotation manual to the statistical analysis of the quantitative data derived from the manually-performed annotations. By presenting this method, the article pursues the twofold purpose of (i) providing a practical tool that can facilitate more reliable, replicable, and transparent analyses, and (ii) fostering a discussion of the best practices that should be observed when manually annotating APPRAISAL.

Keywords: reliability, replicability, transparency, inter-coder agreement, intra-coder agreement, challenges in analyzing APPRAISAL

1 Introduction

APPRAISAL (Martin and White, 2005) has gained increasing recognition as a useful framework for analyzing evaluative phenomena in discourse. Within this framework, manual text annotation has become a popular method for examining and comparing the use of evaluative language resources across texts and corpora (e.g. Bednarek, 2008; Carretero and Taboada, 2014; Don, 2007; Fuoli, 2012; Fuoli and Hommerberg, 2015; Hommerberg and Don, 2015; Lipovsky, 2008, 2011, 2013; Mackay and Parkinson, 2009; O'Donnell, 2014; Pounds, 2010, 2011; Ryshina-Pankova, 2014; Santamaría-García, 2014; Taboada and Carretero, 2012; Taboada et al., 2014). Manual annotation facilitates comprehensive and detailed analyses that would not be possible with purely automatic techniques, given the complex and context-dependent nature of evaluation in discourse (Fuoli and Hommerberg, 2015). But manual annotation may also be seen as an important part of the process of theory building. By applying the set of categories included in the APPRAISAL framework to the annotation of concrete instances of text, we obtain information that can be used to progressively develop and refine the model.

However, annotating APPRAISAL poses a number of challenges, which may hinder the reliability and replicability of analyses. First and foremost, identifying expressions of AP-

PRAISAL in text is a complex and highly subjective task. Evaluative meanings may be conveyed both explicitly and implicitly through an open-ended range of diverse linguistic forms. Moreover, the genre and the communicative context in which a text is produced and consumed have a major impact on our interpretation of the meanings expressed and, as a result, on the annotation process. Thus, analysts need not only to rely on their own intuitions as to what stretches of text appear to communicate some kind of direct or implied stance, but also regarding the appropriate interpretation to be given to expressions in a given context. Classifying evaluative expressions into the categories provided by the model is also a difficult and subjective task. In many cases, multiple interpretations for textual items are possible and the boundaries between the categories are not always clear-cut. In addition, the general-purpose architecture of the framework often clashes with the "contextual specificity of evaluation" (Macken-Horarik and Isaac, 2014: 70). As a consequence, analysts are frequently confronted with the problem of dealing with infelicitous matches between the definitions and examples provided in the literature and the instances found in their texts. Finally, the practicalities of annotating APPRAISAL have not been sufficiently discussed and, to date, there is no well-established, standardized annotation protocol.[1] The lack of a well-defined and widely accepted methodology represents an obstacle for both novice practitioners and experienced analysts, and poses a challenge to achieving transparent and replicable analyses.

How should we deal with the problem of subjectivity? How should we account for our decisions so that our analyses are transparent, reliable, and maximally replicable? What steps are involved in the process of manually annotating text based on the APPRAISAL framework? While there is a growing awareness of methodological issues (e.g. Hommerberg and Don, 2015; Macken-Horarik and Isaac, 2014; Thompson, 2014), the literature on APPRAISAL has provided incomplete answers to these questions. This article seeks to address these challenges and propose some solutions to overcome them. It describes a step-wise method for the manual annotation of APPRAISAL in text and corpora that is designed to optimize the reliability, replicability, and transparency of the analysis. By presenting this technique, the article pursues the twofold purpose of (i) providing a practical tool that can facilitate more transparent, reliable, and replicable APPRAISAL analyses, and (ii) fostering a discussion of the best practices and optimal protocols to be observed when manually annotating APPRAISAL in text. In this sense, this article may be seen as a step towards a formalized manual annotation methodology for APPRAISAL analysis. But the method proposed here has broader potential applications. Given that the issues with which analysts are often confronted when annotating APPRAISAL are common to most types of semantically-oriented analysis, it may be adapted to a wide range of semantic/functional annotation tasks.

The method presented here has been developed as part of a larger project that investigates the discursive negotiation of *trust* in corporate communication (Fuoli and Hommerberg, 2015; Fuoli and Paradis, 2014). It draws on ideas and practices from the fields of *natural language processing* (Pustejovsky and Stubbs, 2012) and *content analysis* (Krippendorff, 2004), where manual text annotation is a widely-used technique. It was employed and tested in a case-study analysis of a specialized corpus of CEO letters, which is presented in Fuoli and Hommerberg (2015).[2] Most of the examples discussed in this article are taken from that corpus and from a larger corpus of corporate social responsibility reports (for information about this text type, see Fuoli, 2012). Several examples from a general corpus of English (COCA) and other sources are also used. The complete list of example sources is provided in the Appendix.

The article is organized as follows. Section 2 reviews some of the main challenges involved in annotating APPRAISAL and explains how these can negatively affect the reliability, replicability, and transparency of APPRAISAL analyses. Section 3 provides an overview of the step-wise annotation method and discusses the solutions implemented to address these problems. The article concludes by assessing the strengths and weaknesses of the proposed approach.

2 Challenges in annotating APPRAISAL

This section reviews some of the most significant challenges that arise in the process of manually annotating text based on the APPRAISAL model. Issues concerning the tasks of (i) identifying expressions of APPRAISAL and (ii) classifying them according to the APPRAISAL typology are discussed in turn. Next, I examine how these challenges can potentially compromise the reliability, replicability, and transparency of APPRAISAL analyses. The discussion of these issues serves as the background and rationale for the method proposed in this article, which is described in section 3. Due to space constraints, a detailed account of the APPRAISAL model is not offered here. A complete overview can be found in Martin and White (2005).

2.1 Challenges in identifying APPRAISAL

Identifying expressions of APPRAISAL, i.e. units of text that perform an evaluative function, is a particularly complex task. Indeed, as Mauranen and Bondi (2003: 269) remark,

> [e]valuation in discourse is an elusive concept. As readers and writers, we seem to be vaguely aware of evaluation being constructed in texts we encounter and produce; it is harder to tell exactly how this happens, that is, which linguistic means are involved, and which (if any) are not.

Given the pervasiveness and apparent elusiveness of evaluation in discourse, it is not surprising that, as Hunston (2004: 158) notes, “many writers on the topic avoid the issue of identification altogether and focus on classifying and analysing instances of evaluation and other aspects of interpersonal meaning”. Hunston’s observation seems to apply well to the case of APPRAISAL theory, where the process of identification of evaluative expressions has not been sufficiently problematized. Martin and White (2005) devote considerable space to describing the framework and presenting various analyses and worked-out examples. However, most of the coding choices made in the analyses are treated as self-evident and unproblematic. But identifying and coding evaluation in text is, in fact, a problematic task for a number of reasons.

First of all, evaluation may be realized through an open-ended range of expressions of varying length and complexity and belonging to any word class, as in (1).

(1) excels [VB]
outstanding [ADJ]
extremely talented [ADJP]
in an enviably strong position [ADV]
one of the world's great enterprises [Complex NP]

Therefore, it is impossible to compile a complete, definitive list of evaluative forms to be searched for in a text or corpus (Hunston, 2011: 13). Accordingly, it is the analyst's ultimate responsibility to decide what counts as evaluation in any given text, which is an inherently subjective process.

The task of identifying APPRAISAL is further complicated by the fact that evaluation is a highly context-dependent phenomenon (e.g. Bednarek, 2006: 8; Hunston, 2011: 10; Macken-Horarik and Isaac, 2014; Martin and White, 2005; Paradis et al., 2012; Thompson and Alba-Juez, 2014). Some expressions may carry an evaluative meaning in certain contexts and co-textual environments, but not in others. Seemingly neutral and descriptive adjectives such as *thin* and *light* are sometimes used for instance in advertising discourse to positively evaluate products, as shown in example (2) (see also Paradis et al., 2012).

(2) There's thin and light. Then there's thin and light on a whole new level. iPod touch has a super-thin aluminum body that feels barely there in your hand or pocket.

If we accept evaluation to be subjective, comparative, and value-laden (Thompson and Hunston, 2000: 13), we can interpret these adjectives as evaluative in this context. They are subjective, in the sense that thinness and lightness are both subjective and relative measures (Paradis, 2001). Their comparative nature is explicitly foregrounded in this context, where the thinness and lightness of the *iPod* is contrasted with the norm (*There's thin and light*). They are value-laden, as they are clearly used to positively evaluate the advertised product, based on the implicit assumption that thinner and lighter phones are better, more desirable phones. The adjective *new* may also be seen to perform an evaluative function in this context, as it is often the case in advertising discourse (Bednarek, 2014). As these examples show, then, the identification of evaluative expressions "cannot be carried out without close attention to contextual factors" (Page, 2003: 213). But this adds subjectivity to the analysis, as the analyst needs to draw on their own knowledge and interpretation of the context in which a text operates, and different analysts may not share the same background information and assumptions (Hommerberg and Don, 2015).

Subjective decisions are not only involved in determining what expressions serve an evaluative function in a certain text, but also in setting their textual boundaries, a process referred to as *unitizing* (Artstein and Poesio, 2008: 580–583; Krippendorf, 2004). For example, should *thin* and *light* in the second sentence of (2) be coded as two separate units? Or should we rather annotate the whole phrase *thin and light on a whole new level* as one single unit? In certain cases, unitization choices can make a significant difference in the analysis, especially if quantitative data are derived from the annotations. Take (3) as an example.

(3) We are well-positioned to generate shareholder value with distinct competitive advantages and a steadfast commitment to the highest standards of ethics, safety, and corporate citizenship.

There are different and equally plausible ways in which evaluation may be unitized in (3). The expressions *steadfast commitment* and *highest standards* may be annotated as separate units, or combined into one single unit. The words *ethics*, *safety*, and *corporate citizenship* may be annotated as three units of, say, JUDGEMENT: PROPRIETY,[3] or as just one. Alternatively, the entire phrase *a steadfast commitment to the highest standards of ethics, safety, and corporate citizenship* may be coded as one, self-contained APPRAISAL unit. Depending on our unitization choices, our analysis could yield between one and five instances.

A related question is how we should treat evaluative expressions joined by a coordinating conjunction (cf. Taboada and Carretero, 2012; Taboada et al., 2014). If we take (4) as an example, should the evaluative adjectives *systematic* and *unwavering* be annotated as two independent evaluative expressions or combined into one?

(4) ExxonMobil's success is underpinned by our commitment to integrity – our systematic and unwavering focus on safety, operational excellence, financial discipline, and high ethical standards.

The implications of such a choice are considerable, given that the observed frequency of certain annotated types could substantially increase if the 'separation rule' was applied to a text that included several such expressions.

Another issue with unitizing is how to deal with *discontinuous* evaluative expressions, i.e. evaluative text spans that are interrupted by either non-evaluative lexical items or by expressions that instantiate a different APPRAISAL category. The following example taken from Carretero and Taboada (2014: 228) shows an expression of GRADUATION (*the most I have ever heard*) interrupted by an intervening expression of ATTITUDE (*boring*) and by a non-evaluative word (*book*).

(5) When I stopped reading my husband laughed and said 'That is the most boring book I have ever read'.

In their study, Carretero and Taboada (2014) opt for the inclusion of the intervening ATTITUDE expression and the non-evaluative word *book* in the GRADUATION span because the software program they use to annotate the corpus, i.e. the UAM Corpus Tool (O'Donnell, 2008), does not currently support the coding of discontinuous text spans (for a comparison of two freely available annotation tools, see section 3.2.2). While this solution is sub-optimal because it introduces redundancy and 'noise' in the coding, separating the two components of the GRADUATION expression would have resulted in the misleading duplication of the annotated instances, with *the most* and *I have ever heard* recorded as two independent items. Clearly then, as this example shows, decisions concerning how to handle discontinuous evaluative expressions may have implications for how evaluation in text is quantified.

Additional identification challenges arise from the distinction in the model between *inscribed* APPRAISAL, i.e. expressions that carry an explicit evaluative charge, and *invoked* APPRAISAL, i.e. seemingly neutral wordings that imply or invite a positive or negative evaluation. While intuitively appealing, this distinction raises several issues. Distinguishing between inscribed and invoked evaluation in text is far from straightforward. As seen above, even apparently descriptive and neutral terms may perform an evaluative function in certain contexts. This entails that there is no simple rule that can be consistently applied to discern the two types. Most importantly, it is unclear how the degree of explicitness of an expression should be determined in the first place. This issue is not sufficiently problematized in Martin and White (2005). While the authors acknowledge that the identification of invoked instances introduces an element of subjectivity into the analysis (Martin and White, 2005: 62), the identification of inscribed instances is presented as essentially unproblematic and the underlying decision-making process as self-explanatory. Yet, when analyzing texts, one is often confronted with ambiguous instances that could be classified equally well as inscribed or invoked. One solution to this problem is to conceive of the distinction between inscribed and invoked evaluation as a cline rather than a dichotomy (Bednarek, 2006: 31). However, while this certainly better reflects the dynamics of meaning-making in discourse, it does not facilitate the task of annotating text, which by its very nature requires either-or distinctions to be made. In sum, identifying *both* inscribed and invoked evaluation is a highly subjective task. Considerations about context can guide our analysis, but, as noted above, different analysts may not share the same information and assumptions, potentially leading to different interpretations.

One further problem concerning the distinction between inscribed and invoked evaluation is what Thompson (2014) calls the 'Russian doll syndrome'. This refers to the cases where an evaluative inscription instantiating one category can be interpreted to indirectly invoke other evaluations in a recursive manner. Thus, for example, the word *superior* in (6), which is taken from Thompson (2014: 59), may be seen as an explicit instance of positive APPRECIATION of the target *magnetic screening.* As Thompson (2014: 59) observes, however, "in the context of a brochure whose main function is to show the company in a good light, this can all be seen as functioning as a token invoking JUDGEMENT of 'we' who 'offer' this superior product".

(6) We offer superior magnetic screening with the minimum of outgassing.

While (6) is a relatively simple example, there can be far more complex cases, where a single evaluative expression may be seen to trigger multiple invoked evaluations at different levels (Thompson, 2014: 59–64). Accounting for these dynamics adds a great deal of complexity to the analysis. Clearly, the more layers are included, the more complex and prone to subjective interpretation it becomes, posing substantial challenges to achieving consistent and reliable annotation.

While the present discussion of identification challenges has focused on the APPRAISAL system of ATTITUDE, the other categories present similar problems. As far as ENGAGEMENT is concerned, for example, the definitions and examples provided in the literature are often insufficient for discerning relevant and non-relevant instances. Negation, for example, is regarded as a key ENGAGEMENT device and features a dedicated category within the system, i.e. DISCLAIM: DENY. The inclusion of negation under the ENGAGEMENT system is justified on the grounds that "[f]rom the dialogistic perspective, negation is a resource for introducing the alternative positive position into the dialogue, and hence acknowledging it, so as to reject it" (Martin and White, 2005: 118). However, not all uses of negation seem to perform this dialogic/intersubjective function. In certain cases, especially when facts, rather than opinions are at stake, negation appears to perform an 'objective/descriptive' function, i.e. to simply reverse the polarity of a statement, rather than to reject an alternative viewpoint. This difference is illustrated in (7a) and (7b) below, which exemplify the 'objective' and 'intersubjective' uses of negation, respectively.

(7) a. For all practical purposes, wartime Washington was a segregated city. There were two school systems, separate and unequal. Restaurants for whites did not admit blacks, although blacks managed, or even owned, some white establishments.
b. 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.

The distinction between objective and inter-subjective uses of negation has not, to the best of my knowledge, been discussed in the literature, and there are currently no set criteria for excluding false positives from the analysis. Thus, the analyst's subjective judgment plays a crucial role in this case too.

In sum, identifying expressions of APPRAISAL in text is a highly complex and subjective task. Section 3 discusses some methodological solutions that can be adopted to deal with these problems. The next section reviews some of the main challenges involved in classifying expressions of APPRAISAL.

2.2 Challenges in classifying APPRAISAL

Just as in the case of identification, the task of classifying evaluative expressions based on the categories provided by the APPRAISAL model poses several conceptual and methodological challenges. First of all, different interpretations for an expression are often equally plausible, and multiple category labels valid. The more fine-grained the analysis is, the more problematic and subjective classification choices become (Read and Carroll, 2012). The adjective *diligent* in (8), for example, would seem to match the JUDGEMENT sub-categories of ABILITY, PROPRIETY, or TENACITY equally well, or at least a justification for classifying it as belonging to each of these categories could be given.

(8) We are a self-sustaining and competitive international, integrated energy company with diligent financial management, strong operating expertise, and an intense focus on optimizing the value of our portfolio.

If we interpret the word *diligent* in (8) in the sense of 'meticulous' or 'persevering', then we should tag this item as an instance of TENACITY. If we take it as highlighting the company's care and conscientiousness in their work, then it could be labeled as PROPRIETY. But the word *diligent* in this context may also be seen to indicate that the company is competently managed, thus foregrounding ABILITY. Neighboring words such as *competitive* and *expertise* could be taken to support this reading.

Albeit less frequently, this type of ambiguity can be found across the 'root' categories of the framework as well. Martin and White (2005: 60), for example, note that certain words, such as *guilty*, *embarrassed*, *proud* or *jealous*, simultaneously construe AFFECT and JUDGEMENT. There are also expressions that can be seen to encode both AFFECT and ENGAGEMENT. One such word is *confident*, which can be interpreted to express both AFFECT: SECURITY (Martin and White, 2005: 51; Bednarek, 2008: 173) and ENGAGEMENT: PROCLAIM. Take (9) as an example.

(9) I am confident that ExxonMobil's competitive advantages position us well to meet these challenges.

One strategy to cope with this type of ambiguities is to allow for double or multiple coding (Macken-Horarik and Isaac, 2014: 88). Rather than annotating expressions with one single category label, we can, when necessary, apply two or more. However, there are several drawbacks to this approach. Most notably, (i) the annotation process becomes more complicated and time consuming, and consistency harder to achieve; (ii) the degree of subjectivity involved in the annotation process grows substantially, as the number of possible choices for each item increases; (iii) quantitative data may be skewed and lose informative value, as the one-to-one correspondence between linguistic expressions and categorial labels would be lost. Therefore, while double coding may be a viable solution to the problem of ambiguous items, we need to be aware of the methodological implications it carries.

Further, certain expressions do not seem to fit comfortably in any of the available categories. This is the case, for example, with generic evaluations such as *good*, *bad* or *great*, which are "semantically underspecified" (Bednarek, 2009: 174) and thus difficult to classify (see also Ben-Aaron, 2005). In other cases, no good match can be found because the evaluative meaning of certain expressions is highly context-specific (Hommerberg and Don, 2015; Macken-Horarik and Isaac, 2014). Hommerberg and Don (2015), for example, observe that in wine reviews, words such as *explosive*, *closed*, or *ageworthy* are commonly used to assess the qualities of wines. The evaluative meaning of these expressions, i.e. the type of values that they foreground, is not adequately captured by any of the categories in the APPRAISAL framework. Macken-Horarik and Isaac (2014) observe similar gaps in the context of narrative discourse. The strategy adopted in both these studies is that of modifying and expanding the original model to fit the specific discursive context under study. Hommerberg and Don (2015), for example, add several categories to the system of APPRECIATION, e.g. INTENSITY, MATURITY and DURABILITY. Adapting the APPRAISAL framework to the specificities of the discourse type at hand can be an effective way of achieving more accurate and informative analyses. This choice, however, poses some methodological challenges. For example, annotators who are not familiar with the genre may not possess the specialized knowledge that is necessary to successfully apply the new categories. In fact, Hommerberg and Don (2015) report substantial inter-coder *dis*-agreement over some of the newly introduced categories. One of the possible causes that the authors identify for the low level of agreement observed precisely lies in the lack of field-specific expertise by one of the annotators.

One additional problem is that the distinction between the categories in the framework is not always clear-cut. As far as the system of ATTITUDE is concerned, for example, the boundary between JUDGEMENT and APPRECIATION is in some cases blurred (e.g. Bednarek, 2009; Ben-Aaron, 2005; Martin and Rose, 2003; Martin and White, 2005; Thompson, 2014). Ambiguity arises when qualities that are normally attributed to people are ascribed to the outcome of their behavior instead. Thompson (2014: 57) discusses the following example from a film review.

(10) But what they've got – and what we've got – is a distinctive, demanding, deeply intelligent picture from a first-class director.

The expression *first-class* in (10) can be unproblematically tagged as an instance of JUDGEMENT; it serves in this context to evaluate a human being, i.e. the prototypical target of JUDGEMENT expressions, and can be used to highlight positive qualities of both inanimate things (e.g. *a first-class hotel*) and people. Classifying the phrase *intelligent picture* is less straightforward. Although what is ostensibly evaluated is the movie, this expression clearly implies a positive assessment of the director as well. His intellectual qualities are transferred to the product of his work, resulting in a 'mismatch' between the actual target of the assessment and the value evoked by the evaluative term *intelligent* (Thompson, 2014: 56–59). Should this phrase then be coded as an instance of JUDGEMENT OR APPRECIATION?

According to Bednarek (2009: 167), there are two criteria by which expressions of ATTITUDE can be classified, i.e. (i) according to the type of lexis used and (ii) according to the entity that is evaluated. In cases such as (10), where *judging lexis* is used to evaluate a non-human target, precedence can either be given to the lexis or the target as the primary classification criterion, although both aspects should be taken into account for a complete analysis (Bednarek, 2009: 184). Thompson (2014: 58) suggests to take "the Target at face value", and classify all instances where a non-human target is evaluated as APPRECIATION, even when judging lexis is used. This applies to all non-human targets, including nominalizations, on the grounds that "[a] nominalization is a non-human entity (albeit a virtual entity); and it is therefore a target of APPRECIATION (even if judging lexis is used)" (Thompson, 2014: 58). But one potential problem with this approach is that in many cases the evaluative expression would not easily fit into any of the subcategories of APPRECIATION. Take the examples of the words *industry-leading* and *disciplined* in (11).

(11) ExxonMobil Chemical has delivered industry-leading performance through disciplined implementation of strategies that have been proven over numerous business cycles.

None of the categories of APPRECIATION could easily accommodate these expressions, so it is not clear how these items should be handled at more delicate levels of analysis, beyond the root category of APPRECIATION. On the other hand, it is clear that their function in this context is to positively evaluate the target *ExxonMobil Chemical* (interpreted as a human entity), so annotating them as APPRECIATION would seem to make the analysis unnecessarily convoluted. Ultimately, however, whichever criterion is used, this decision should be made explicit, and it should be consistently applied throughout the analysis to optimize transparency and reliability (see section 2.3). Yet, as Bednarek (2009: 167) notes, this type of information is rarely disclosed in studies based on the APPRAISAL framework.

One issue that has received limited attention in the literature, but which has both important theoretical and practical implications, relates to how we should deal with evaluative expressions that are not used to communicate an actual positive or negative evaluation of some target, but rather to refer to a hypothetical or *irrealis* scenario. (12) illustrates the distinction between 'actual' and 'irrealis' evaluations.

(12) a. We have been playing a leading role in carbon capture and storage.
b. We *aim* to play a leading role in the growing low-carbon energy sector.

While *a leading role* is used in (12a) to afford an actual positive evaluation of the target *we*, the same expression identifies a desirable future state of affairs in (12b). This shift is triggered in this example by the verb *aim*, but other verbs, such as *intend*, *want*, *aspire*, verbs of belief, modals of possibility, and the future auxiliary *will* also have the ability to activate similar irrealis scenarios. Clearly, the discourse function of actual and irrealis evaluative expressions is different; (12b) may be seen to invoke a positive evaluation of the target *we*, rather than inscribe it, even though the expression *a leading role* is clearly explicitly evaluative. This difference should be accounted for in the analysis if we aim to capture and correctly represent the discursive function of evaluative expressions. This is particularly important when working with discourse types such as corporate reports, which are replete with forecasts and statements of intents like (12b). However, no clear guidelines for dealing with this type of co-textual dynamics are currently available. Bednarek (2008) uses the label *hypotheticality* to refer to similar phenomena, but she does not discuss this issue in detail. The APPRAISAL framework draws a distinction between *realis* and *irrealis* AFFECT, where the latter concerns feelings related to future states, e.g. desires and fears (Martin, 2000: 150; Martin and White, 2005: 48). However, this distinction is different from the one discussed here. In fact, even when a speaker expresses a desire or fear toward a yet unrealized situation (e.g. *I want to be a millionaire*), he or she still communicates an actual feeling, just as the writer in (12a) expresses an actual positive self-evaluation.

While the discussion of ambiguities in the APPRAISAL literature has tended to focus on the system of ATTITUDE, classification challenges involve ENGAGEMENT as well. The boundaries between dialogic *expansion* and *contraction* are not always clear, especially when it comes to the sub-categories of ENTERTAIN (with expansion) and PROCLAIM (with contraction). The former groups "those wordings by which the authorial voice indicates that its position is but one of a number of possible positions and thereby [...] makes dialogic space for those possibilities" (Martin and White, 2005: 140). The latter includes expressions that "act to limit the scope of dialogistic alternatives in the ongoing colloquy" (Martin and White, 2005: 121). Certain markers of ENGAGEMENT may be interpreted as instances of ENTERTAIN in certain contexts, but of PROCLAIM in others. The epistemic verb *believe*, for instance, is listed under ENTERTAIN in Martin and White (2005: 98). In many contexts, as for example in (13), *believe* clearly performs a dialogically expanding function, as defined above.

(13) When will ordinary people get to go into space? I believe that flights will become available when private companies realize that it can become a profitable tourism enterprise.

There are cases, however, in which this verb seems to perform a contractive function instead. One such case is when *believe* is modified by instensifiers such as *firmly*, as in (14).

(14) We firmly believe deepwater drilling can be done safely and in an environmentally sensitive manner.

In (14), *believe* appears to behave in a similar way as other instances of PROCLAIM: PRONOUNCE (e.g. *I contend that*, *it is absolutely clear to me that*); it serves to represent the proposition as highly warrantable and thus to suppress or rule out alternative positions (Martin and White, 2005: 98). The use of the plural personal pronoun *we* may be seen to add to the contractive effect of *believe* in this context, as the proposition is not merely represented as one individual's personal point of view, but as fully endorsed by a group of people. In other cases, e.g. when *believe* co-occurs with other dialogically contractive markers, it is unclear which category would better capture its context-specific function. Consider, for example, (15).

(15) The year' s return was comparable to that of the Standard & Poor's (S&P) 500 index, but lagged that of our peers – a performance that we believe does not reflect our company' s potential.

In fact, disagreement between independent annotators was found in the classification of instances such as (15) in the process of annotating the corpus used for the study presented in Fuoli and Hommerberg (2015). Martin and White (2005: 103–104) recognize that the function of ENGAGEMENT expressions "may vary systematically under the influence of different co-textual conditions, and across registers, genres and discourse domains". These conditions, however, are not discussed in any detail, thus classifying instances of *believe* and other expressions of ENGAGEMENT is often a very subjective exercise.

In conclusion, as Macken-Horarik and Isaac (2014: 81) remark, evaluation "resists enclosure in analytical boxes and frustrates the 'either-or' distinctions that are central to the [APPRAISAL] system network". Yet, exclusive choices are an inescapable part of the process of text annotation. The number and variety of highly subjective decisions that, as discussed above, are involved in both the task of identifying and classifying expressions of APPRAISAL may represent a challenge to achieving acceptable standards of reliability, replicability, and transparency. These issues are discussed in more detail in the next section.

2.3 Reliability, replicability, transparency

As the discussion above has shown, annotating APPRAISAL involves a number of highly subjective and, in some cases, rather arbitrary choices at different levels and stages of the process. In addition, it is a complex and cognitively demanding task, as evaluation is a pervasive feature of discourse, and annotation choices are not always straightforward. These issues pose a serious challenge to the reliability, replicability, and transparency of analyses based on the model. These three issues are fundamentally intertwined, as explained below.

In general, reliability can be defined as "the extent to which a measurement procedure yields the same answer however and whenever it is carried out" (Kirk and Miller, 1986: 19). There are three main types of reliability: (i) *test-retest* reliability, (ii) *internal consistency* reliability, and (iii) *interrater* reliability (Cozby and Bates, 2011: 96–101). Test-retest reliability, also referred to as *stability*, is "the extent to which a measuring or coding procedure yields the same results on repeated trials" (Krippendorff, 2004: 215). In the context of manual corpus annotation, it concerns analysts' ability to accurately reproduce their own annotations on different occasions, separated by a time interval. Achieving this type of reliability when applying the APPRAISAL model to the annotation of text is a difficult task for various reasons. If analysts solely rely on their own intuitions, rather than on explicit and formalized annotation criteria, their knowledge and understanding of the annotation principles may change over time, leading to inconsistent coding choices. The fuzziness of the model is also a major obstacle to achieving stability. Since analysts are often confronted with the possibility of multiple, equally valid choices, it may be hard for them to maintain consistency across coding sessions.

Internal consistency, in the context of behavioral research, refers to the consistency of results across test items that are designed to measure the same variable (Cozby and Bates, 2011: 99). As far as manual text annotation is concerned, it measures the extent to which an annotator is consistent in applying the coding guidelines, treating similar textual items in the same way throughout a text or corpus. Achieving adequate internal consistency may be difficult in the case of APPRAISAL analysis. First of all, just as they may hinder stability, the identification and classification problems discussed above may negatively affect internal consistency too. The lack of clarity about identification criteria and the ambiguity that characterizes some category definitions may lead analysts to unwittingly treat similar items in a different way throughout a text. Second, the process of manual annotation is a highly cognitively demanding task, and consistency can be negatively affected by a variety of psychological and contingent factors, such as memory constraints, fatigue, and fluctuating concentration and motivation levels. This may lead, for example, to false negatives (i.e. unintentionally missed items), and to classification incongruities within the same text or produced during the same session. Furthermore, as noted above, some expressions may perform an evaluative function in certain contexts and co-textual environments only. This means that, with the exception of very 'stable' expressions, annotation choices necessarily need to be made on a case-by-case basis. Clearly, this adds to the cognitive effort required to carry out the task. But it also implies that even *assessing* internal consistency is not straightforward, and may be impossible to carry out automatically.

Interrater reliability measures "the degree to which a process can be replicated by different analysts" (Krippendorff, 2004: 215). In the context of manual corpus annotation, it is a function of the extent to which independent coders working under the same guidelines agree on the categories assigned to text units. A high level of interrater agreement indicates that the annotation task is well defined and thus potentially replicable. Conversely, a low level of agreement indicates either that the coding scheme is defective or that the raters need to be retrained (Artstein and Poesio, 2008; Pustejovsky and Stubbs, 2012). Achieving robust interrater reliability is a necessary step to ensure the replicability of the annotation procedure and of the analysis results. It is also a prerequisite for claiming the *validity* of the annotation guidelines, i.e. that they adequately capture the phenomenon under study (Artstein and Poesio, 2008: 557; Krippendorf, 2004: ch. 11). In fact, as Artstein and Poesio (2008: 557) observe, "just as in real life, the fact that witnesses to an event disagree with each other makes it difficult for third parties to know what actually happened". However, achieving a satisfactory level of interrater reliability is, in the case of APPRAISAL, a very difficult task. APPRAISAL analysis is, as demonstrated above, highly subjective; the analyst's individual knowledge, beliefs and reading position will inevitably affect the annotation process. In addition, the model's often under-specified or ambiguous category definitions leave much room for subjective and even arbitrary decisions.

One way in which reliability and replicability can be substantially improved is by defining explicit annotation criteria. As Kirk and Miller (1986: 41) put it, "reliability depends essentially on explicitly described observational procedures". As shown above, annotating APPRAISAL entails multiple decisions at different levels, e.g. concerning unitization criteria, whether multiple coding is allowed, or about the principles that should guide the classification of expressions. In addition, in many cases the definitions and examples included in Martin and White (2005) alone do not provide a sufficiently clear and precise basis for producing reliable text annotations. Context-specific definitions and guidelines are in most cases necessary to implement the model in concrete text analyses, and adapt it to the specificities of the linguistic data under study and the goals of the project. All the decisions made and the context-specific annotation guidelines formulated during the annotation process should be explicitly formulated and made available to other analysts. This is not only crucial for improving reliability and replicability, but also for ensuring transparency, i.e. allowing others to trace and fully understand the annotation process and correctly interpret the results. If the principles that have guided the annotation process are not made explicit and available, it is impossible to correctly and critically interpret and assess the results of an analysis. Moreover, by disclosing the annotation criteria, we enable other researchers to contribute to their improvement, and, ultimately, to a progressive and collaborative development of the APPRAISAL model.

Despite their obvious importance, it is unfortunate that the issues of reliability, replicability, and transparency have received little attention in the literature based on the APPRAISAL framework. The question of reliability has been framed by Martin and White (2005: 207) in terms of an opposition between subjectivity and objectivity, and dismissed on the grounds that "so-called objectivity is impossible". It is still rare for empirical studies using the framework to report measures of inter-coder or intra-coder agreement (see section 3.2.4), and the issue of reliability is rarely directly addressed (for exceptions, see Fuoli, 2012; Fuoli and Hommerberg, 2015; Hommerberg and Don, 2015; Read and Carroll, 2012; Ryshina-Pankova, 2014; Taboada and Carretero, 2012). In many cases, quantitative findings are presented, but only limited information is given about the specific criteria adopted in the annotation of the corpus (cf. the numerous issues and choices discussed above), or about how this process was carried out (e.g. Don, 2007; Lipovsky, 2011; Mackay and Parkinson, 2009; Marshall et al., 2010; Pounds, 2011; Santamaría-García, 2014). Ultimately, however, if these methodological issues are inadequately attended to, no finding or conclusion can be accurately interpreted or questioned, making APPRAISAL analysis more similar to an "idiosyncratic and impressionistic commentary on discourse – somewhat like a literary exegesis – rather than a replicable linguistic analysis" (Thompson, 2014: 64). In the next section, I describe a step-wise method for annotating APPRAISAL that offers some practical solutions to these challenges, and tools to optimize reliability, replicability, and transparency.

3 The step-wise annotation method

This section presents a step-by-step method for manually annotating APPRAISAL that is designed to optimize reliability, replicability, and transparency, while retaining the flexibility of the model. It draws upon methodological insights from natural language processing (Pustejovsky and Stubbs, 2012) and content analysis (Krippendorff, 2004), two areas where manual text annotation is an established and widely-used technique. First, I explain the basic principles underlying the method. Next, I describe the key steps it involves.

3.1 General principles

The discussion above has shown that annotating APPRAISAL implies a number of choices, ranging from general issues, such as what categories should be considered and what the optimal level of granularity should be, to the more practical questions of how ambiguous items should be treated and instances unitized. In order to optimize transparency and replicability, we need to record and disclose all the choices we make. Thus, the first, basic principle behind the method described here is the following.

PRINCIPLE 1. All choices should be accounted for.

The inherent subjectivity of the annotation process may lead us to think that reliability is unattainable. However, in those cases where, for example, interrater reliability has been measured, the results show substantial agreement between independent coders, at least for the higher-order categories of the model (Fuoli, 2012; Fuoli and Hommerberg, 2015; Read and Carroll, 2012; Taboada and Carretero, 2012). Note that this does not imply that perfect reliability can actually be reached, nor that this should be our ultimate goal. As Martin and White (2005: 161–164) point out, every text represents a meaning potential and is susceptible to different readings. Thus, a certain amount of individual variation in the identification and classification of evaluative expressions is unavoidable. Rather, what we should aim for is *maximum* reliability; the annotation guidelines should be progressively developed and refined until reliability does not show any sign of further improvement. This will allow us to achieve the most transparent, robust, and replicable analysis possible. Thus, the second general principle underlying the present method is the following.

> PRINCIPLE 2. The annotation guidelines should be tested and refined until maximum reliability is achieved.

Reliability should always be assessed and reliability scores reported. This will enhance transparency and will provide a useful basis for interpreting the results. Where reliability is low, the potential reasons for this should be explicitly discussed, as cases of disputed annotations may indicate the need for improving the model and provide useful insights for how to do so (see e.g. Hommerberg and Don, 2015). In view of this, the third fundamental principle on which this method is built is the following.

> PRINCIPLE 3. Reliability should always be assessed, and reliability scores reported and discussed.

These fundamental principles have been implemented into a seven-step annotation procedure, which is described in the next section.

3.2 An outline of the steps

Figure 1 provides an overview of the step-wise method. The steps are sequential, with a loop between steps 4 and 5, as shown by the feedback arrow. The arrow indicates that, as stated above, the annotation guidelines should be tested and progressively improved until maximum reliability is reached. The procedure for doing this is described in detail below. In the following sections, I discuss each step in detail and provide practical implementation guidelines.

3.2.1 Step 1: Define the scope of the project

Not all APPRAISAL analyses necessarily need to cover the whole range of categories of the model. The *annotation scheme*, i.e. the list of categories and subcategories to be used in the annotation of the corpus, should be defined in relation to the research question(s) at hand, the goals of the study, and the specific features of the texts under study. Thus, the

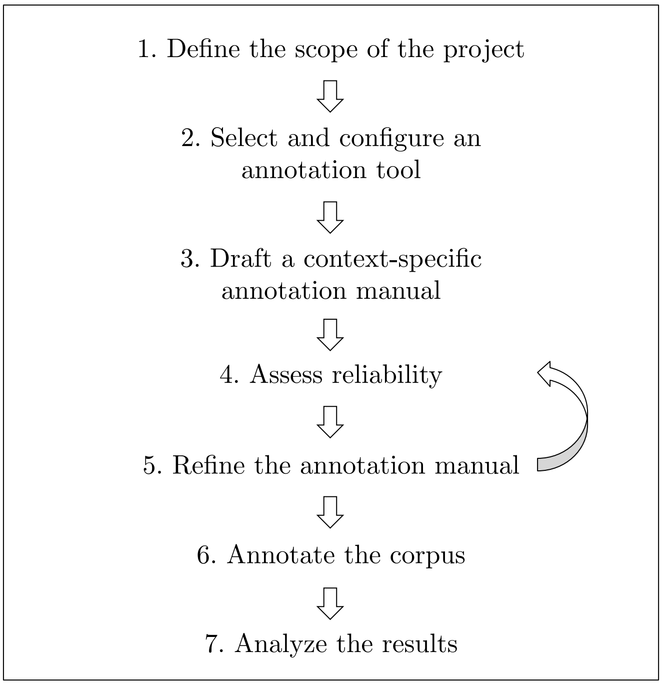
1. Define the scope of the project
2. Select and configure an annotation tool
3. Draft a context-specific annotation manual
4. Assess reliability
5. Refine the annotation manual
6. Annotate the corpus
7. Analyze the results

Figure 1: An overview of the step-wise method for annotating APPRAISAL

first step consists in determining the scope of the annotation project, by selecting the categories that are potentially relevant and that should be included in the scheme. The scheme will be progressively developed and refined throughout the annotation process.

To create a preliminary version of the annotation scheme, random samples from the corpus may be informally annotated by simply manually underlining instances of the categories that are envisaged to be significant, given the research question(s). At the same time, any unexpectedly frequent or interesting phenomena should also be noted. Based on this initial exploratory analysis, categories from the draft coding scheme may be added or removed as required. The appropriate level of granularity for the coding scheme should be determined based on the goals of the project, the time and resources available, and the specificities of the texts under study. For example, if the preliminary informal exploration of the corpus indicates that instances of a certain category are particularly frequent, it might be useful to allow for a higher degree of granularity for that category. Note, however, that the more fine-grained the analysis is, the more time consuming and complex the annotation process will be.

3.2.2 Step 2: Select and configure an annotation tool

A practical and effective way to annotate corpora is to use a dedicated corpus annotation tool. Using an annotation tool, rather than a simple text editor or word processor, can help make the annotation process faster and more systematic. In addition, quantitative data can be generated accurately and automatically.

Several freeware programs are available (for a review, see O'Donnell, 2014). The most widely used by researchers working with the APPRAISAL framework is the UAM CorpusTool (O'Donnell, 2008, 2012). UAM provides a user-friendly interface for annotating text based on a coding scheme defined by the user. It also incorporates various useful text analysis tools, including a statistics module, which allows the researcher to automatically compute the frequency of annotated categories. In addition, UAM features an 'autocoding' function, which allows users to automatically assign a given label to all instances of a certain word or expression in the corpus. This function can be very useful for improving internal consistency and expedite the annotation process.

An alternative to UAM is the CAT annotation tool (Bartalesi Lenzi et al., 2012). CAT offers similar functions to the UAM tool, with some advantages and limitations. One advantage that CAT has over UAM is that it allows the annotation of discontinuous text spans (see section 2.1). Further, CAT is a web-based tool, which means that researchers can access and work on the same annotation project from different locations. This can greatly simplify the task of discussing and reconciling disagreements between independent annotators (see section 3.2.4), as the project files are simultaneously accessible and changes instantly available to all participants in the project. Finally, CAT has an in-built inter-coder agreement tool that automatically computes inter-coder agreement scores between independent analysts.[4] As for its limitations, unlike the UAM corpus tool, CAT does not currently incorporate any auto-coding function. Also, CAT's interface is not equipped to display the hierarchy between the categories in the coding scheme, which may be helpful when working with the APPRAISAL framework, in particular when the granularity of the annotation scheme is very fine.

Every annotation tool has strengths and weaknesses that need to be considered when choosing the one to be used in a given project. Most importantly, however, we should choose the annotation tool that we feel most comfortable using. It is important to stress that the choice of tool should have no impact on the results or on the reliability of the annotation process, provided that the annotation guidelines are explicit and transparent.

3.2.3 Step 3: Draft a context-specific annotation manual

As discussed above, one crucial step for optimizing reliability, replicability, and transparency is to account for all decisions and provide explicit and detailed guidelines that other researchers can review and use (Principle 1). These guidelines should be incorporated into an *annotation manual*, a document that should (minimally) include an outline of the annotation scheme, the category definitions, explicit rules for applying the definitions to the data set under study, including detailed instructions on how to unitize instances of APPRAISAL and deal with ambiguous or multifunctional items, and illustrative examples. The manual should be *context-specific*; the definitions and coding guidelines should be shaped around the specific characteristics of the texts to be annotated. This means that any relevant consideration about context or about the discourse type under study that could affect the interpretation and annotation of evaluative expressions (e.g. what the main communicative purpose is, who the intended audience is) should be made explicit, to the extent possible. Ideally, the manual should be included in any publication stemming from the annotation project for which it was developed. If this is not possible, it should be made available to other researchers upon request. An example of a context-specific annotation manual can be found in the appendix to Fuoli and Hommerberg (2015). Taboada et al. (2014) also provide very detailed and clearly-formulated coding guidelines that may be used as a model for the creation of an annotation manual.

To test the robustness and clarity of the draft annotation manual before moving on to the next stage, a random sample from the corpus may be annotated, and the guidelines edited until they appear to be robust enough, that is, until they seem to cover most cases and leave as little room as possible for ambiguity. If a large number of instances that do not match any of the categories included in the APPRAISAL model are found, context-specific or 'ad-hoc' categories may be created. In that case, the new categories should be given a transparent label, and accompanied by a clear definition and explicit annotation instructions. Any specific rules for dealing with ambiguous items, examples and exceptions that seem necessary at this stage should be added to the manual, so as to make it as clear, transparent, and comprehensive as possible.

3.2.4 Step 4: Assess reliability

As discussed in section 2.3, assessing and optimizing reliability is necessary to ensure that the data derived from the annotations are consistent and trustworthy, and that the analysis is maximally replicable. As seen above, there are three types of reliability, i.e. stability, internal consistency, and interrater reliability. At this stage, the stability and interrater reliability of the annotation procedure will be tested. The test results will then be used to improve and refine the annotation guidelines at step 5. In most cases, the guidelines will need to be revised multiple times before a satisfactory level of reliability is reached. Steps 4 and 5 should thus be repeated as many times as necessary, until maximum reliability is achieved (Principle 2). This is shown in Figure 1 by the feedback arrow connecting steps 4 and 5.

The existence of a loop between steps 4 and 5 implies that the development of the annotation guidelines is a dynamic and iterative process (Pustejovsky and Stubbs, 2012). Pustejovsky and Stubbs (2012) refer to this process as the "Model-Annotate-Model-Annotate", or *MAMA* cycle. The MAMA cycle comprises four steps: model, annotate, evaluate, and revise (Pustejovsky and Stubbs, 2012: 28). Simply put, the annotation guidelines developed on the basis of theory are applied to the annotation of a text sample by two or more annotators. Their annotations are then compared and interrater reliability assessed. If reliability is high enough, the coding of the corpus may proceed. Conversely, if agreement is low, the guidelines are revised, improved, and tested again. The same procedure is implemented here. In addition to interrater reliability, however, also stability is assessed.

Stability is measured by means of an *intra-coder agreement* test, interrater reliability by means of an *inter-coder agreement* test. The procedure for conducting these tests is similar. The only difference is that in the former case two annotations made by the same person on different occasions are compared, whereas in the latter the annotations made by two or more independent coders are set against each other. The way in which agreement is calculated depends on the type of task at hand. As seen above, annotating APPRAISAL involves two main tasks, i.e. (i) the identification of APPRAISAL expressions, and (ii) their classification. As far as the identification task is concerned, agreement can be calculated using *precision*, *recall* and *F-measure* scores (Fuoli and Hommerberg, 2015; Read and Carroll, 2012; Taboada and Carretero, 2012). For the classification task, Cohen's chance-corrected *kappa* coefficient can be used (Cohen et al., 1960). Fuoli and Hommerberg (2015) provide a description of these scores and a practical illustration of a procedure that can be followed to gather inter-coder agreement data and calculate the scores. Due to space constraints, this information is not reproduced here. For a detailed discussion and comparison of different inter-coder agreement measures, see Artstein and Poesio (2008).

Stability should be tested first and the guidelines revised (step 5) until maximum intra-coder agreement is achieved. Next, the same procedure should be carried out for interrater reliability. Random samples from the corpus under study should be used for both tests. Where assessing interrater reliability is not possible due to resource or time constraints, stability may be used as the sole criterion for reliability. Clearly, however, stability is a comparatively weaker measure of reliability (Krippendorff, 2004: 215).

3.2.5 Step 5: Refine the annotation manual

At this stage, the results of the stability and interrater reliability tests should be used to improve and refine the annotation guidelines so as to optimize their robustness and clarity. Incongruities between either the two annotations produced by the same annotator or by independent coders may indicate that the relevant definitions and instructions in the annotation manual are not well defined or clear enough, or that additional rules need to be added. The manual should therefore be revised accordingly. Category definitions may be modified, ad-hoc rules and categories added, and additional illustrative examples given, if necessary. A discussion of disagreements with the other annotator(s) may help to identify the weaknesses of the guidelines, and provide useful insights for resolving ambiguities and correct defects. After refining the manual, reliability should be tested again (step 5). As stated above, this procedure should be repeated until a satisfactory and/or stable level of agreement has been reached.

3.2.6 Step 6: Annotate the corpus

When the reliability of the annotation procedure has been optimized, the entire corpus may finally be annotated. This task may be performed by one or more people, provided that interrater reliability is high. The corpus should be annotated following the guidelines included in the manual as strictly as possible. At this stage, care should be taken to ensure that internal consistency is maintained (see section 2.3). Steps should be taken to minimize fatigue and cognitive load, which may negatively affect consistency. Dividing the annotation process into two separate sub-tasks, one involving the identification of APPRAISAL items and the other focusing on their classification, and taking regular breaks may help to optimize focus and minimize strain. Reviewing each text multiple times may also contribute to achieving higher internal consistency.

3.2.7 Step 7: Analyze the results

Once the corpus has been fully annotated, the data may be analyzed using both quantitative and qualitative techniques. Quantitative analysis may be carried by counting and comparing the frequency of APPRAISAL categories across texts, but also using more sophisticated multifactorial techniques (see, e.g. Divjak, 2006; Geeraerts et al., 1994; Glynn, 2009; Gries, 1999). Multifactorial statistics can be used to investigate the effect of multiple textual and contextual factors on linguistic behavior, in this case the use of APPRAISAL expressions. Manual corpus annotation provides an ideal basis for conducting this type of analysis. The factors of interest can easily be included in the annotation scheme, and all APPRAISAL expressions identified can be classified according to specific textual and contextual variables. In a multifactorial analysis of APPRAISAL in a small corpus of corporate social responsibility reports, for example, Fuoli and Glynn (2013) consider as many as fourteen textual and contextual factors, such as hypotheticality, evaluative target, polarity, topic, and reporting company. These factors are used to model the data derived from the annotations and test which of them has the strongest and most significant effect on the choice of APPRAISAL expressions. In addition to frequency counts and multifactorial methods, O'Donnell (2014) describes several statistical techniques that have been specifically conceived for APPRAISAL analysis. For a comprehensive account of statistical methods for linguistics, see e.g. Gries (2013) or Baayen (2008).

4 Conclusion

This article has described a step-wise method for the manual annotation of APPRAISAL that is designed to optimize reliability, replicability, and transparency. The method offers some practical solutions to the problem of subjectivity, which, as the discussion above has shown, may hinder the reliability and replicability of analyses and have a major impact on the results. By creating explicit, detailed, and context-specific annotation guidelines and refining them until maximum reliability is reached, the subjectivity involved in annotating APPRAISAL is not eliminated, but 'tamed' and controlled for, to the maximum extent possible. As discussed above, this will ensure that the data derived from the annotations are consistent and trustworthy, and that the analysis is transparent and maximally replicable. The method presented here also provides a comprehensive and widely applicable procedure for annotating text based on the APPRAISAL model, which is not currently available in the literature. This article thus offers both experienced analysts and novice practitioners a practical tool to produce systematic quantitative and qualitative APPRAISAL analyses.

Clearly, the method presented here has several limitations as well. First of all, it is comparatively more time consuming than a more informal approach to manual annotation, where the issues of reliability, replicability, and transparency are not addressed. The higher cost involved in following this procedure would be offset by the higher quality of the analysis, but time and resource constraints may still make its application simply impracticable.

Further, this method might be better suited for quantitative analysis than for the qualitative, interpretive approach promoted by Macken-Horarik and Isaac (2014) and adopted by many scholars working within the APPRAISAL framework. While there is some degree of overlap between the latter approach and the one advocated for here, for example in the recognition of the importance of context and the need for adapting the APPRAISAL framework to the specificities of the discourse type and context under study, the methodological and epistemological principles underpinning the two methods are substantially different. Macken-Horarik and Isaac (2014) embrace the fuzziness of the APPRAISAL model, rather than seeking to minimize it. The analytical procedure they outline allows much room for subjective and idiosyncratic choices, which is precisely what the technique presented here aims to reduce. Indeed, as Hood (2004: 15) suggests, "there is a trade-off in the choice of approach in any one study". The present method enables more systematic and reliable analyses, but it is less flexible than Macken-Horaric and Isaac's (2014) approach, and possibly less suitable for analyzing invoked APPRAISAL. On the other hand, Macken-Horaric and Isaac's (2014) approach allows for sophisticated and thorough qualitative analyses, but disregards reliability and is more difficult to translate into systematic and structured annotation projects from which trustworthy quantitative data may be derived. But these two approaches are by no means mutually exclusive. Rather, they may be seen as complementary; they may successfully combined to harness their respective strengths and achieve more robust and complete analyses of evaluation in discourse. Given that context is crucial for correctly interpreting and analyzing evaluation, and a thorough and deep knowledge of the discourse type at hand is key to achieving high-quality annotations, qualitative/interpretive analysis may be seen as an important or even necessary first step in the preparation of any annotation project, and as an indispensable source of information for correctly interpreting the patterns uncovered through manual corpus annotation and quantitative analysis.

Finally, another limitation is that general criteria and standard thresholds for reliability have not been established yet for the task of annotating APPRAISAL, so it is unclear when an analysis can be considered reliable *enough*. Artstein and Poesio (2008: 557) argue 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*" (emphasis mine). Clearly, this definition leaves room for discretion. The approach adopted here is that we should always aim at maximum reliability, and that data can be considered maximally reliable when any further change to the guidelines does not translate into any improvement in the reliability scores. But this approach raises the question of when maximum reliability is too low to be considered acceptable. This issue requires further discussion.

Clearly, much work needs to be done to develop a standardized and widely applicable methodology for annotating APPRAISAL in text, as well as to refine the model and make it more robust. I hope that this article will stimulate a productive discussion around these issues.

Acknowledgments

I wish to thank Carita Paradis (Lund University, Sweden) for her support and guidance throughout the preparation of the article. I am also grateful to Charlotte Hommerberg (Linnaeus University, Sweden) for valuable discussions, insight, and fruitful collaboration. Finally, I would like to thank the Editorial Board of Functions of Language and three anonymous reviewers for their constructive comments on the manuscript.

Notes

[1] Some considerations about annotation methodology are offered in O'Donnell (2008, 2012, 2014), where the author demonstrates the application of the UAM corpus tool, a computer program specifically designed to support and facilitate the task of coding text based on the SFL framework, to the task of annotating APPRAISAL. While O'Donnell's work, as discussed below, stands as a major contribution to the literature in that it provides a very useful tool that can help make APPRAISAL analyses more systematic, it does not directly address the issues of reliability, replicability, and transparency in relation to the task of annotating APPRAISAL, which are the primary focus of the current paper. Further, the author does not provide a complete set of guidelines for how to actually conduct manual corpus annotation through all its various phases. O'Donnell's work mainly focuses on how to use the UAM corpus tool, e.g. how to handle text files, how to configure the tool to conduct manual annotation tasks, or how to search texts and visualize the results. The scope of the current paper is broader, and goes beyond giving practical instructions on how to use a tool for manual corpus annotation.

[2] The method is presented here in a more formalized way than in Fuoli and Hommerberg (2015), where some of the steps that are here presented as separated are conflated into macro-steps.

[3] By convention, category labels throughout the paper are indicated in small caps. When two category labels appear next to each other separated by a colon, which indicates that the second term is a subcategory of the first.

[4] It should be noted, however, that the inter-coder agreement score that the tool employs, i.e. the Dice similarity index (Dice, 1945), is different and more conservative than those that have been traditionally used for the task of annotating APPRAISAL (see Fuoli and Hommerberg, 2015; Read and Carroll, 2012; Taboada and Carretero, 2012). This does not mean that the Dice coefficient cannot be used for this task, but task-specific benchmarks should be developed to be able to accurately interpret the results.

References

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Bartalesi Lenzi, V., Moretti, G., and Sprugnoli, R. (2012). CAT: the CELCT Annotation Tool. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S. (Eds.), *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Bednarek, M. (2006). *Evaluation in media discourse: Analysis of a newspaper corpus.* London: Continuum.

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

Bednarek, M. (2009). Language patterns and ATTITUDE. *Functions of language*, 16(2): 165–192.

Bednarek, M. (2014). An astonishing season of destiny! Evaluation in blurbs used for advertising TV series. In Thompson, G. and Alba-Juez, L. (Eds.), *Evaluation in context*, pp. 197–220. Amsterdam and Philadelphia: John Benjamins.

Ben-Aaron, D. (2005). Given and news: Evaluation in newspaper stories about national anniversaries. *Text & Talk*, 25(5): 691–718.

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. et al. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1): 37–46.

Cozby, P. and Bates, S. (2011). *Methods in behavioral research (11th ed.).* New York: McGraw-Hill.

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302.

Divjak, D. (2006). Ways of intending: Delineating and Structuring Near-Synonyms. In Gries, S. T. and Stefanowitsch, A. (Eds.), *Corpora in cognitive linguistics. Corpus-based approaches to syntax and lexis*, pp. 19–56. Berlin: Mouton.

Don, A. (2007). An approach to the analysis of textual identity through profiles of evaluative disposition. In *Proceedings of the Australian Systemic Functional Linguistics Association 2007 Conference*, Sydney, Australia.

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

Fuoli, M. and Glynn, D. (2013). Computer-assisted manual annotation of evaluative language expressions: Bridging discourse and corpus approaches. Paper presented at the Evaluative Language and Corpus Linguistics Workshop, Lancaster University, 22 July 2013. Workshop webpage: http://ucrel.lancs.ac.uk/cl2013/evaluativelanguage.php [Accessed 22 September 2015].

Fuoli, M. and Hommerberg, C. (2015). Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, 10(3): 315–349.

Fuoli, M. and Paradis, C. (2014). A model of trust-repair discourse. *Journal of Pragmatics*, 74: 52–69.

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

Glynn, D. (2009). Polysemy, syntax, and variation. A usage-based method for Cognitive Semantics. In Evans, V. and Pourcel, S. (Eds.), *New directions in cognitive linguistics*, pp. 77–104. Amsterdam: John Benjamins.

Gries, S. T. (1999). Particle movement: A cognitive and functional approach. *Cognitive Linguistics*, 10(2): 105–146.

Gries, S. T. (2013). *Statistics for linguistics with R: A practical introduction.* Berlin & New York: de Gruyter.

Hommerberg, C. and Don, A. (2015). Appraisal and the language of wine appreciation: A critical discussion of the potential of the appraisal framework as a tool to analyse specialised genres. *Functions of Language*, 22(2): 161–191.

Hood, S. (2004). *Appraising Research: Taking a stance in academic writing.* PhD thesis, Faculty of Education. University of Technology, Sydney. http://www.grammatics.com/appraisal/hoodS-phd-links.htm. [Accessed 22 September 2015].

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.

Kirk, J. and Miller, M. L. (1986). *Reliability and validity in qualitative research.* Beverly Hills, CA: Sage.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, CA: Sage.

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

Lipovsky, C. (2011). 'It's really a great presentation!': Appraising candidates in job interviews. *Linguistics & the Human Sciences*, 4(2): 161–185.

Lipovsky, C. (2013). Negotiating one's expertise through appraisal in CVs. *Linguistics and the Human Sciences*, 8(3): 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): 729–753.

Macken-Horarik, M. and Isaac, A. (2014). Appraising Appraisal. In Thompson, G. and Alba-Juez, L. (Eds.), *Evaluation in context*, pp. 67–92. Amsterdam and Philadelphia: John Benjamins.

Marshall, C., Adendorff, R., and de Klerk, V. (2010). The role of APPRAISAL in the NRF Rating System: An analysis of Judgement and Appreciation in peer reviewers' reports. *Southern African Linguistics and Applied Language Studies*, 27(4): 391–412.

Martin, J. (2000). Beyond exchange: Appraisal systems in English. In Hunston, S. and Thompson, G. (Eds.), *Evaluation in text: Authorial stance and the construction of discourse*, pp. 142–175. Oxford: Oxford University Press.

Martin, J. R. and Rose, D. (2003). *Working with discourse: Meaning beyond the clause.* London: Continuum.

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

Mauranen, A. and Bondi, M. (2003). Evaluative language use in academic discourse. *Journal of English for Academic Purposes*, 2(4): 269–271.

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. (2012). Appraisal analysis and the computer. *Revista Canaria de Estudios Ingleses*, 65: 115–130.

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

Page, R. (2003). An analysis of APPRAISAL in childbirth narratives with special consideration of gender and storytelling style. *Text-Interdisciplinary Journal for the Study of Discourse*, 23(2): 211–237.

Paradis, C. (2001). Adjectives and boundedness. *Cognitive Linguistics*, 12(1): 47–64.

Paradis, C., van de Weijer, J., Willners, C., and Lindgren, M. (2012). Evaluative polarity of antonyms. *Lingue e linguaggio*, 11(2): 199–214.

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): 106–137.

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

Pustejovsky, J. and Stubbs, A. (2012). *Natural language annotation for machine learning.* Sebastopol, CA: O'Reilly Media.

Read, J. and Carroll, J. (2012). Annotating expressions of Appraisal in English. *Language Resources and Evaluation*, 46(3): 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*, pp. 281–302. Amsterdam and Philadelphia: John Benjamins.

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*, pp. 387–411. Amsterdam and Philadelphia: John Benjamins.

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 and the Human Sciences*, 6(1-3): 275–295.

Taboada, M., Carretero, M., and Hinnell, J. (2014). Loving and hating the movies in English, German and Spanish. *Languages in Contrast*, 14(1): 127–161.

Thompson, G. (2014). AFFECT and emotion, target-value mismatches, and Russian dolls: Refining the APPRAISAL model. In Thompson, G. and Alba-Juez, L. (Eds.), *Evaluation in context*, pp. 47–66. Amsterdam and Philadelphia: John Benjamins.

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

Thompson, G. and Hunston, S. (2000). Evaluation: An introduction. In Hunston, S. and Thompson, G. (Eds.), *Evaluation in text: Authorial stance and the construction of discourse*, pp. 1–27. Oxford: Oxford University Press.

Appendix: List of example sources

Example (1): adapted from BP's 2008 Annual Review. Available at: http://www.bp.com/en/global/corporate/investors/results-and-reporting/annual-report/annual-reporting-archive.html [last accessed 28 October 2015]

Example (2): Apple Inc. UK Website. Available at: https://www.apple.com/uk/ipod-touch/ [last accessed: 1 October 2014]

Example (3): ExxonMobil 2011 Summary Annual Report, p. 2. Available at: http://ir.exxonmobil.com/phoenix.zhtml?c=115024\&p=irol-reportsAnnual [last accessed 28 October 2015]

Example (4): ExxonMobil 2010 Summary Annual Report, p. 2. Available at: http://ir.exxonmobil.com/phoenix.zhtml?c=115024\&p=irol-reportsAnnual [last accessed 28 October 2015]

Example (5): Carretero and Taboada (2014: 228)

Example (6): Thompson (2014: 59)

Example (7a): Corpus of Contemporary American English (COCA). Accessible online at: http://corpus.byu.edu/coca/

Example (7b): BP Annual Report and Form 20-F 2010, p. 11. Available at: http://www.bp.com/en/global/corporate/investors/results-and-reporting/annual-report/annual-reporting-archive.html [last accessed 28 October 2015]

Example (8): ConocoPhillips 2008 Annual Report, p. 3. Available at: http://www.conocophillips.com/investor-relations/company-reports/Pages/annual-report-archive.aspx [last accessed 28 October 2015]

Example (9): ExxonMobil 2008 Summary Annual Report, p. 3. Available at: https://bib.kuleuven.be/files/ebib/jaarverslagen/EXXONMOBILE_2008.pdf [last accessed 28 October 2015]

Example (10): Thompson (2014: 57)

Example (11): ExxonMobil 2009 Summary Annual Report, p. 32. Available at: https://bib.kuleuven.be/files/ebib/jaarverslagen/EXXONMOBILE_2009.pdf [last accessed 28 October 2015]

Example (12a): BP Sustainability Review 2009, p. 18. Available at: http://www.bp.com/content/dam/bp/pdf/sustainability/group-reports/bp_sustainability_review_2009.pdf [last accessed 28 October 2015]

Example (12b): BP Sustainability Review 2009, p. 16. Available at: http://www.bp.com/content/dam/bp/pdf/sustainability/group-reports/bp_sustainability_review_2009.pdf [last accessed 28 October 2015]

Example (13): Corpus of Contemporary American English (COCA). Accessible online at: http://corpus.byu.edu/coca/

Example (14): ConocoPhillips 2010 Summary Annual Report, p. 4. Available at: http://www.conocophillips.com/investor-relations/company-reports/Documents/SMID_394_IR_CompanyReports_AR_Archive_2010_English.pdf [last accessed 28 October 2015]

Example (15): ConocoPhillips 2008 Annual Report, p. 5. Available at: http://www.conocophillips.com/investor-relations/company-reports/Pages/annual-report-archive.aspx [last accessed 28 October 2015]