Lang Resources & Evaluation (2012) 46:421–447
DOI 10.1007/s10579-010-9135-7

ORIGINAL PAPER

Annotating expressions of Appraisal in English

Jonathon Read · John Carroll

Published online: 12 December 2010

Abstract In the context of Systemic Functional Linguistics, Appraisal is a theory describing the types of language utilised in communicating emotion and opinion. Robust automatic analyses of Appraisal could contribute in a number of ways to computational sentiment analysis by: distinguishing various types of evaluation, for example affect, ethics or aesthetics; discriminating between an author's opinions and the opinions of authors referenced by the author and determining the strength of evaluations. This paper reviews the typology described by Appraisal, presents a methodology for annotating Appraisal, and the use of this to annotate a corpus of book reviews. It discusses an inter-annotator agreement study, and considers instances of systematic disagreement that indicate areas in which Appraisal may be refined or clarified. Although the annotation task is difficult, there are many instances where the annotators agree; these are used to create a gold-standard corpus for future experimentation with Appraisal.

Keywords Appraisal · Corpus annotation · Inter-annotator agreement · Opinion · Subjectivity · Systemic Functional Linguistics

1 Introduction

The increasingly rapid development of the World Wide Web has facilitated the dissemination of opinion on a scale greater than ever before, not only from traditional publishers but also the general public. The pieces published by news

J. Read (✉)
Department of Informatics, University of Oslo, PO Box 1080, Blindern, 0316 Oslo, Norway
e-mail: jread@ifi.uio.no

J. Carroll
School of Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, UK
e-mail: j.a.carroll@sussex.ac.uk

agencies, papers and broadcasters are often reproduced online, while blogging technologies enable Web users to readily post their thoughts and experiences. There are many professional and enthusiast review websites, and online retailers often encourage their customers to review their purchases. Indeed, this abundance of product reviews has motivated the creation of opinion-aggregation websites such as Metacritic.com.

This wealth of readily-available opinion has spurred much research into the automatic analysis of opinion-bearing text. For instance, Wiebe et al. (2004) investigated features of text that indicate whether a proposition is objective or subjective (that is, if it is information believed to be factual by the individual, or if it represents an opinion held by the individual). Others have sought to classify text according to its sentiment (Pang et al. 2002; Turney 2002): assuming it is opinion-bearing, is the opinion positive or negative about its subject? Some researchers have carried out classification according to several dimensions, seeking to identify different types of emotion (Subasic and Huettner 2001). Other studies have conducted deeper analyses that determine facets of opinion-bearing expressions such as the holder, target, and nature (Wiebe et al. 2003).

There exist several frameworks from various fields of academic study, such as cognitive science, linguistics and psychology, that can inform and augment analyses of sentiment and opinion. Ekman (1993), for instance, derived a list of six basic emotions from subjects' facial expressions which Strapparava and Mihalcea (2007) employed as classes in an affect recognition task. Hyland (1998) described the linguistic phenomenon of hedging, where writers express the degree to which an opinion is speculative or unconfirmed. Di Marco and Mercer (2004) used features based on hedging to determine the nature of relationships in scientific articles. Gratch and Marsella (2004) developed a cognitive model of appraisal that considered several variables affecting the strength of appraisal, such as the relevance and urgency of an event, and the degree to which the ego is involved. This model was created for use by avatars simulating an emotional reaction, but could be used to inform analyses of opinion if suitable indicators of these variables could be found. Wiebe et al. (2005) created a scheme for the annotation of the mental and emotional state conveyed by text. Their scheme distinguished between explicit expressions (such as *the US fears a spill-over*) and subjective expressive elements where the affective state is implied by words that contain negative connotations (e.g. *we foresaw electoral fraud but not daylight robbery*).

In this article we focus on Appraisal (Martin and White 2005), a theory of evaluative language developed by researchers working in Systemic Functional Linguistics.[1] The theory distinguishes between types of attitude (personal affect, judgement of people and appreciation of objects), and describes how authors use language to communicate their engagement with other writers, and to amplify or diminish the strength of their attitudes and engagements. Texts annotated with these aspects of language could potentially enhance existing computational techniques for

[1] Note the distinction between the Systemic Functional Linguistic theory of Appraisal (Martin and White 2005) discussed in this article, and the Cognitive Psychology theory of Appraisal (Scherer et al. 2001), which deals with how emotions are affected by assessment of events

sentiment, opinion and affect analysis by considering the type and strength of evaluation communicated, and identifying when and how authors report the opinions of others.

There are currently no machine-readable Appraisal-annotated texts publicly available. Aspects of the theory have been demonstrated using examples from genres as different as news reporting (White 2002; Martin 2004) and poetry (Martin and White 2005). As syntactic constructions and lexical choices are likely to vary greatly across such genres, it is inappropriate to quantitatively examine such examples. Instead this article presents a quantitative study across several documents in the same genre addressing a number of important issues including areas of difficulty in the annotation task and inter-annotator agreement. The study has the additional benefit of creating a machine-readable corpus annotated with Appraisal types for further research by the Appraisal, Corpus Linguistics and Computational Linguistics communities.

This article is structured as follows: Sect. 2 reviews the three subsystems of Appraisal (Attitude, Engagement and Graduation) and considers other computational explorations of the theory. Section 3 describes the challenges presented by the Appraisal annotation study and the methodology employed during its course. Section 4 details how inter-annotator agreement was measured by analogy with scores used to evaluate information extraction systems, and considers instances of systematic disagreement. One example of disagreement is explored in further detail in Sect. 5, which presents the results of a sentence-based annotation exercise conducted using several annotators. Section 6 describes how the annotations from the main study were compiled into a gold-standard, and Sect. 7 presents conclusions.

2 Appraisal

Appraisal,[2] summarised by the systems network[3] in Fig. 1, is a Systemic Functional Linguistic theory of evaluation in text (Martin and White 2005). It consists of three subsystems that operate interactively: Attitude is concerned with one's personal feelings (emotional reactions, judgements of people and appreciations of objects); Engagement considers the positioning of oneself with respect to the opinions of others (heterogloss) and with respect to one's own opinions (monogloss); while Graduation addresses how language functions to amplify or diminish the attitude and engagement conveyed by a text. The theory describes a typology of words that not only covers emotions and opinions but also the manner in which authors engage with their audience and other authors, and how authors modify the strength of opinions expressed.

[2] Typographical note: throughout this article the labels of classes in the Appraisal theory are indicated using Small Capitals.

[3] Systems networks are Systemic Functional Linguistic tools that display the relations between features in a theory. The features serve as entry points into subsequent systems. Square brackets indicate a logical *or* relationship, while a logical *and* relationship is depicted by angle brackets.

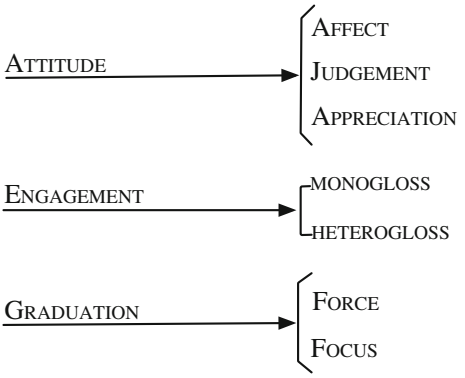

Fig. 1 A systems network depicting the structure of Appraisal resources. *Square brackets* indicate a logical *or* relationship, while a logical *and* relationship is depicted by angle brackets

2.1 Attitude: emotion, ethics and aesthetics

The subsystem of ATTITUDE is a framework for three areas of personal feeling: emotion, ethics and aesthetics. The hierarchy is depicted in Fig. 2. All types of attitude can also be analysed according to their Polarity, be it positive or negative.

2.1.1 Affect

Descriptions of personal emotion are referred to as AFFECT. The Appraisal system considers four subclasses of affect: INCLINATION is concerned with items that express some degree of personal desire towards or against phenomena (e.g. *miss, long for, yearn* versus *wary, fearful, terrorised*[4]); terms of HAPPINESS deal with internal mood (e.g. *cheerful, like, jubilant* versus *sad, dejected, joyless*); one's environmental and social well-being is covered by SECURITY (e.g. *confident, assured, trusting* versus *uneasy, anxious, startled*); and one can also express SATISFACTION with one's goals (e.g. *pleased, thrilled, involved* versus *jaded, angry, bored*).

2.1.2 Judgement

Evaluations of people (JUDGEMENTS) are divided into two types: ESTEEM and SANCTION. Judgements of esteem consist of evaluations of NORMALITY (a person's behaviour compared with what a culture considers normal, e.g. *lucky, normal, fashionable* versus *unluck, odd, dated*), CAPACITY (the capability of a person, e.g. *powerful, witty, successful* versus *mild, dull, unsuccessful*) and TENACITY (the dependability of a person, e.g. *plucky, reliable, faithful* versus *timid, unreliable, unfaithful*). Judgements of sanction are to do with VERACITY (the honesty of a

[4] These examples were first presented by Martin and White (2005), as were all the examples that appear in this section.

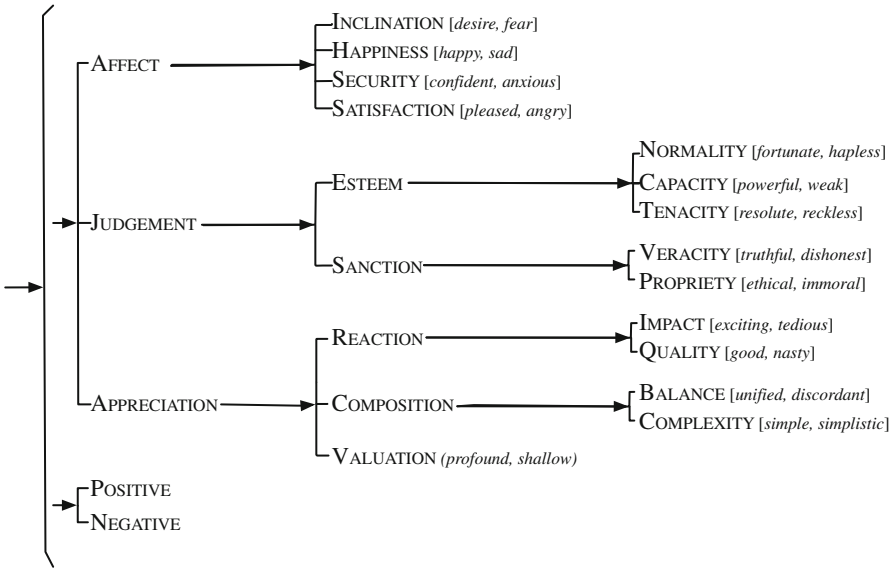

Fig. 2 The ATTITUDE subsystem

person, e.g. *truthful, frank, discrete* versus *dishonest, deceptive, blunt*) or PROPRIETY, e.g. *good, fair, polite* versus *bad, unfair, rude* (how well a person's ethics match those of the culture).

2.1.3 Appreciation

Communication of aesthetic evaluations are instances of APPRECIATION, which is concerned with the different ways we evaluate all things, including man-made objects, performances and natural phenomena. Appreciations are classified as either REACTIONS, assessments of COMPOSITION, or VALUATION of the thing in question. The three types of appreciation may be thought of as levels in a cline of sophistication: reaction being instinctive appreciation, composition being perceptive appreciation and valuation being cognitive appreciation.

Reactions are with respect to the thing's IMPACT (e.g. *engaging, exciting, lively* versus *tedious, ascetic, dull*) or QUALITY (e.g. *good, lovely, welcome* versus *nasty, plain, off-putting*), whereas assessment of composition is concerned with BALANCE (e.g. *unified, shapely, consistent* versus *discordant, flawed, uneven*) or COMPLEXITY (e.g. *simple, precise* versus *simplistic, wooly*). VALUATION describes the worth of something (e.g. *profound, creative, priceless* versus *shallow, everyday, pricey*), but Martin and White (2005) point out that the instances of this class are often dependent on the field of discourse (affected by aspects such as its participants, process and circumstance) because the supposed value of a thing is variable from register to register.

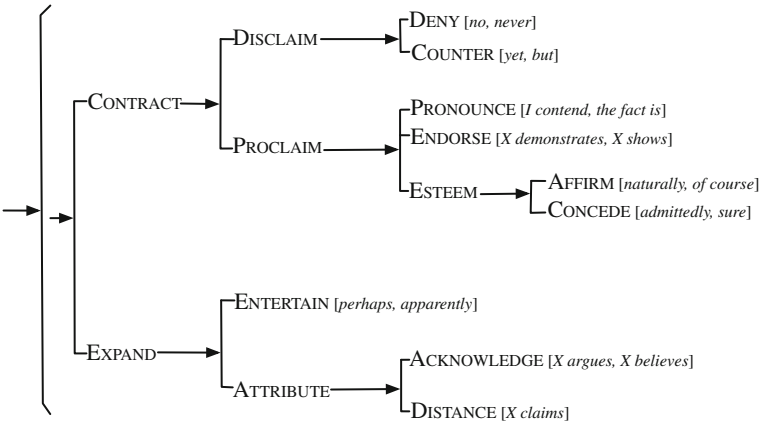

Fig. 3 The ENGAGEMENT subsystem

2.2 Engagement: appraisals of appraisals

Through ENGAGEMENT, Appraisal addresses the linguistic phenomena by which authors construe their point of view and the resources used to adopt stances towards other authors' perspectives. This assumes that all text conveys opinion to some degree and that all writing represents both explicit and implicit responses to other opinions. Furthermore, these responses can be either retrospective (responding to previously expressed opinions) or prospective (anticipating audience response and including counter-responses).

The resources of engagement are depicted as a systems network in Fig. 3. The taxonomy enables a classification of the particular type of dialogistic positioning associated with meanings, and allows one to describe the differences afforded by the various meanings. In this system utterances are said to be either monoglossic or heteroglossic. Monoglossic text does not allow for any viewpoints other than the author's as it contains bare assertions, whereas heteroglossic text allows for two or more viewpoints and their relationships to be represented. Heteroglossic text can CONTRACT or EXPAND dialogue.

2.2.1 Dialogic expansion

Dialogic expansions make allowances for the stances of others, thus opening up more points of view for discussion. Dialogue can be expanded through the entertainment or attribution of propositions.

When the authorial voice accepts that there are other valid positions other than its own it ENTERTAINS these alternatives. This can be realised through: modal auxiliaries (*may, might, could, must*,); modal attributes (*it's possible that, it's likely that*); constructions such as *in my view*; and cognitive reports (*I suspect that, I doubt that*). Such locutions are often interpreted as markers of author confidence, such as in the

literature discussing hedging (Hyland 1998). However, Martin and White (2005) argue that when viewed dialogistically, they also connote a heteroglossic environment in which the author recognises alternatives existing in the current social context. Writers can also entertain the position of others through evidential means/language (such as *seems, apparently* and *suggests* and rhetorical questions).

Martin and White (2005) analyse linguistic phenomena that dissociate propositions from the author and assign them to others as ATTRIBUTION. Typically attribution is realised through reporting (e.g. *X said, Y believes*). Note that there exists a degree of overlap in the lexical items of entertainment and of attribution, although these instances are easily disambiguated by the subject of the construct (e.g. *I believe* versus *they believe*). A proposition is attributed through either ACKNOWLEDGEMENTS or DISTANCES.

An author ACKNOWLEDGES a position when they cite some other author's viewpoint but do not explicitly indicate their own stance. In this case reporting verbs tend to be employed (e.g. *say, report, state, declare, announce, believe* or *think*). Acknowledgements in isolation facilitate a façade of impartial citation. This is useful in some registers (news reporting, for instance), but in other genres where impartiality is unnecessary, author alignment can be conveyed through adverbs (e.g. *X rightly observes* versus *Y foolishly predicts*).

In contrast, an author can overtly DISTANCE themselves from a reported proposition. This is realised through a subset of the reporting verbs (e.g. *claims*). While unmodified acknowledgements remain fairly ambiguous with regards to solidarity, distancing attributions clearly state the author's alignment with respect to the extra-textual proposition; the author explicitly denies any responsibility for the position.

2.2.2 *Dialogic contraction*

Dialogic contractions challenge the position of others, reducing the range of alternative viewpoints through expressions that either DISCLAIM or PROCLAIM.

DISCLAIM covers constructions that invoke an alternative point of view in order to reject it. One subtype of this construction, DENY, occurs when a writer explicitly denies another's viewpoint through negation (e.g. *no, not, nothing, never*). An alternative point of view is acknowledged and rejected, clearly disaligning the author with the explicit or implicit position holder. A second kind of disclamation is that of COUNTERing, where the author responds to a presupposition with a contrary statement (e.g. *Even though we are getting divorced, Bruce and I are still best friends*). This is often conveyed though conjunctions and connectives (e.g. *although, however, yet*, and *but*). It can also be realised though certain adverbials that act as marks of counter-expectation (*surprisingly*, for instance).

In contrast to rejecting some contrary position, when an author PROCLAIMS they in some other way seek to limit the set of options for responses by other authors in an ongoing heteroglossic dialogue. An author who overtly declares their positive alignment with a proposition CONCURS with that proposition. This is usually marked with lexical items such as *of course, naturally, unsurprisingly* and *certainly*. Also,

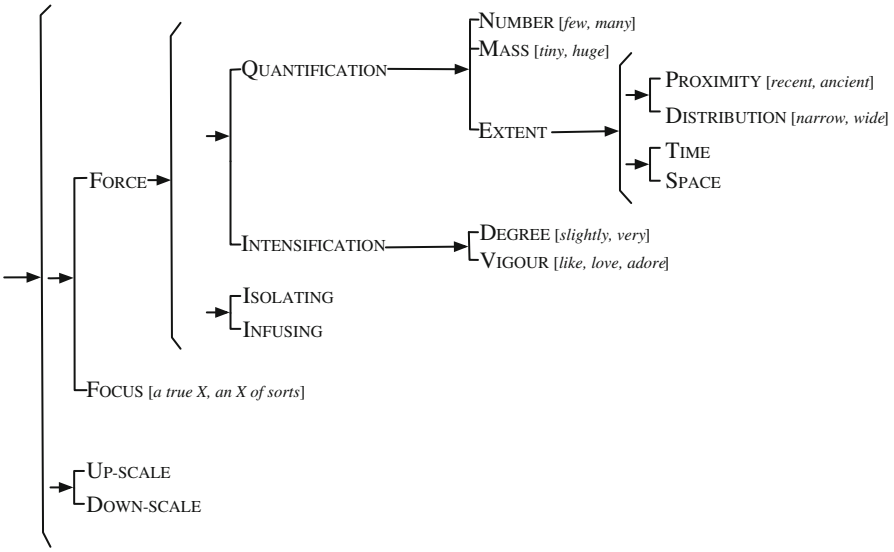

Fig. 4 The GRADUATION subsystem

as with ENTERTAIN, certain types of rhetorical questions will indicate proclamation depending on both the linguistic and real-world context. Concurring can either be in terms of AFFIRM (e.g. *obviously*) or CONCEDE (e.g. *admittedly*). Under the class of ENDORSE, Martin and White (2005) refer to formulations that attribute propositions to external sources and frame these propositions as "maximally warrantable", that is, the author strongly endorses the value of the proposition. Proclamations of the PRONOUNCE type include expressions that encode emphases which indicate an author's position (e.g. *I contend, the fact is, the truth is, we can conclude, you must agree* and clausal intensifiers such as *really* and *indeed*).

2.3 Graduation: strength of evaluation

Martin and White (2005) consider the resources by which writers alter the strength of their appraisal as a system of GRADUATION, summarised by the systems network in Fig. 4. Gradability is a general property of both attitude and engagement. Graduation in attitude enables authors to convey greater or lesser degrees of positivity or negativity, while in engagement graduation scales authors' conviction in their propositions.

2.3.1 Focus

The subsystem of FOCUS considers the graduation of semantic categories that are not typically considered as scalable (e.g. *they don't play real jazz* or *they play jazz, sort of*). Normal experiential perspective tells us that someone either plays jazz or they do not, but in both of these examples the writer maps an evaluative expression

that marginalises the performance; membership of the set of those who play jazz is no longer true or false but fuzzy. Focus can either SHARPEN (amplify) or SOFTEN (diminish). Sharpening formulations have also been labelled as 'intensifiers', 'boosters' and 'amplifiers' (Labov 1984; Hyland 2000).

2.3.2 Force

The subsystem of FORCE alters assessment in terms of intensities and quantities.

Formulations of INTENSIFICATION operate on qualities (e.g. *slightly foolish*, *very foolish*), on processes (*slightly hindered us*, *greatly hindered us*) and on modalities (*it's just possible, it's very possible*). Intensification can be realised grammatically through isolated items such as the examples given above (and including maximising words such as *utterly, totally* and *completely*), through repetition (*laughed and laughed and laughed*), or through figurative metaphors (*ice cold* and *crystal clear*). Intensification may also be realised lexically through infused items. This term refers to instances where the intensification is in the manner of lexical choice rather than modifying constructions. For example, in *this [disquieted | startled | frightened | terrified] me*, the degree of intensification of fear conveyed relies on cultural norms regarding the lexical choices.

QUANTIFICATION constructions scale attitudes with regard to amount and extent, in terms of: NUMBER (*few, many*), MASS (*small, large*) and EXTENT in space and time with respect to either PROXIMITY (*near, far; recent, ancient*) or DISTRIBUTION (*sparse, wide-spread; short-term, long-term*). As with intensifying constructions, quantifiers can operate through isolation or infusion. Examples of infusing lexical items with respect to size include *he's a mountain of a man* in contrast to *she's a slip of a girl*.

2.4 Computational uses of Appraisal

Taboada and Grieve (2004) reported probably the first computational experiment with Appraisal Theory, using some of its insights in a system for document-level sentiment classification. Document sentiment was determined in terms of a binary classification (positive versus negative) by applying Turney's (2002) SO-PMI-IR method on extracted adjectives. For each adjective they estimated a 'Potential' value for affect, judgement and appreciation using a method similar to SO-PMI-IR, calculating the mutual information between the adjective and three pronoun-copular pairs: *I was* (Affect); *he was* (Judgement) and *it was* (Appreciation). While the pronoun-copular pairs seem at first glance to be compelling markers of the respective subsystems, they are somewhat unsatisfactory. For example, they constrain affect to be limited to what is experienced by oneself, whereas affect in Appraisal includes descriptions of others' emotional states. We can expect a high degree of intersection between the different sets obtained from retrieval queries using these pairs (e.g. *I was a happy X, he was a happy X, it was a happy X*).

Whitelaw et al. (2005) argued that 'Appraisal Groups' should be the atomic units when using machine learning techniques for sentiment analysis. Their Appraisal

Groups were loosely based on Appraisal Theory in that they derive a frame of sentiment comprised of:

Attitude:	affect | judgement | appreciation
Orientation:	positive | negative
Force:	low | neutral | high
Focus:	low | neutral | high
Polarity:	marked | unmarked

Note that in Whitelaw et al.'s paper, *polarity* referred to whether an item is negated (marked) or otherwise (unmarked). Typically in the sentiment analysis literature, polarity refers to the positivity or negativity of text (which Whitelaw et al. called *Orientation*). Their process began with a semi-automatically constructed lexicon of these appraisal groups. The lexicon was expanded from seed terms taken from Martin and White's (2005) book, and supplemented with modifiers that change the force, focus and polarity. The appraisal group features supplemented bag of words machine learning techniques for sentiment analysis, resulting in modest gains in accuracy.

Argamon et al. (2007) considered how Appraisal-inspired lexicons might be automatically constructed. In particular, they created a lexicon with entries labelled with Attitude type (Affect, Appreciation or Judgement) and Force (low, median, high or maximum). They employed Esuli and Sebastiani's (2005) method of expanding classes of words by training on the aggregated WordNet glosses of seed terms (also taken from Martin and White's (2005) examples). Argamon et al. evaluated the accuracy of Naïve Bayes and Support Vector Machine classifiers trained in this way by attempting to label words in a manually constructed lexicon (built by expanding the seed set using manually-verified entries in two thesauruses). They found that the Naïve Bayes classifier performed best overall with an F_1 of 0.345 for attitude types (baseline 0.155) and 0.352 for force (baseline 0.239). Support Vector Machines, however, achieved better precision.

The experiments reviewed above are interesting contributions to sentiment analysis research inspired by aspects of Appraisal Theory. However, these aspects are arbitrarily selected based on the researchers' intuitions about what might benefit sentiment analysis and, to date, no work has investigated the impacts of the engagement subsystem on sentiment analysis tasks. Furthermore, recognition of the range of attitude-bearing types may have implications for sentiment analysis in terms of differences in domains (for example, financial newswire text might focus on judgement of companies, whereas a movie review could contain the author's affective reaction to the movie). Finally, Appraisal analysis is an interesting task in itself. For instance, identifying expressions of judgement relating to an organisation could be useful for brand reputation analysis, while the automatic identification of text evaluating attributes of products could be important for opinion-mining with respect to consumer satisfaction.

3 Annotation methodology

The effort required to collect data to support tasks in sentiment analysis depends to a large extent on the complexity of the task for which it is gathered. For example, much research in this area has concentrated on document-level classification. Product reviews are an appealing source of data for such tasks, since many web sites allow reviewers to augment their unstructured text with a quantitatively expressed sentiment rating. Downloading these texts and extracting the rating provides a large number of labelled documents (Pang et al. 2002; Turney 2002). Acquiring even grater numbers of automatically labelled texts is possible with other indicators of positivity and negativity. Read (2005) constructed a corpus of messages labelled with sentiment from Usenet posts, by assuming that a 'smile' emoticon indicated positive text, while a 'frown' emoticon flagged negative text. This particular approached proved to be unreliable, though, as the data collected contained a great deal of noise, indicating that emoticons are not definite denotations of sentiment. This technique is nevertheless appealing as it enables a large amount of data to be collected, and has been successful applied in studies of emotion-bearing language (Yang et al. 2007).

One might consider applying annotations describing aspects of Appraisal to Wiebe et al's (2005) MPQA corpus as the two schemes are complementary, in that Wiebe et al.'s scheme considers the manner of private state expression, whereas Appraisal considers the different types of private state. However, as the MPQA corpus was sourced from newswire articles, it contains comparatively few expressions of Affect, and so we sought to collate a new corpus that adequately represented all types described by Appraisal.

The Appraisal annotation exercise described in this article was conducted on a corpus of unedited and complete book reviews. Book reviews are appealing for this study as they are likely to contain examples of each of Appraisal's many types. One may find descriptions of characters' emotions, judgement of author proficiency, appreciation of the qualities of books, and engagement with the opinions expressed by the authors of the books under review. We obtained the articles from the websites of four British newspapers (The Guardian, The Independent, The Telegraph and The Times). Samples were taken on two different dates (31 July 2006 and 11 September 2006). Each review author is represented only once in the corpus, but articles are often introduced with a paragraph written by an unnamed editor. The corpus is comprised of 38 documents, containing a total of 36,997 words.

There are a number of possible approaches to obtaining human annotations. Read (2004) attempted automatic labelling of sentences according to several classes in a psychological model of affect. To evaluate this task, text was collected from a website thought to be likely to contain a high number of propositions involving affect. A web application allowed human annotators to ascribe a class to each sentence. The task was open to any person who cared to take part, and annotators could annotate as many or as few sentences as they desired. This approach has the advantage that it allows for a large number of annotations from multiple judges. However, on particularly complex tasks, such as this, the approach suffers in that most coders will be unfamiliar with the model of affect and thus are likely to make

misinterpretations. On the other hand, Mihalcea and Chklovski (2003) demonstrated that naïve annotators' contributions can be valuable in the selection of word sense annotations, and ensured quality by acquiring tags from two annotators per item. Other researchers have employed trained annotators (Wiebe et al. 2005), or a combination of trained and naïve annotators (Bruce and Wiebe 1999).

Two human annotators were employed in this Appraisal annotation exercise (*d* and *j*), annotating text independently. The annotators were not given specific instructions as both were familiar with the literature concerning Appraisal Theory (as summarised in Sect. 2). Their instructions were to annotate Appraisal-bearing terms with the Appraisal type (one of 32 types) presumed to be the intention of the author, and also to assign a Polarity (positive or negative) to attitude annotations and a Scaling (up or down) to graduation annotations. The judges employed a custom-developed tool to annotate the documents that was designed according to the exact level of functionality for this task.[5] Annotations were made by selecting a length of text, and clicking a button corresponding to an Attitude, Engagement or Graduation annotation, which in turn displayed a panel of radio buttons listing the possible options for the annotation type. Annotations were held in a modifiable list, and indicated in the text panel using colour-coded highlighting.

We considered a range of alternative annotation strategies. The first of these allowed only a single token per annotation. However, in many instances a unit of Appraisal spans multiple words:

Example 1 The design was deceptively$^{\text{Veracity}}$ simple.$^{\text{Complexity}}$

Example 2 The design was deceptively simple.$^{\text{Complexity}}$

Example 1 shows an analysis of a sentence using single tokens, which incorrectly indicates that the sentence includes a judgement of someone's honesty, whereas Example 2 gives the correct analysis, that it is an appreciation of a design. This example demonstrates that it is necessary to annotate larger units of Appraisal-bearing language than single tokens.

Annotating multi-word expressions, however, increases the complexity of the annotation task, and reduces the likelihood of agreement between the judges, as the annotated tokens of one judge may be a subset of, or overlap with, those of another. We therefore experimented with constraining judges by asking them to tag entire sentences only. This resulted in other problems since there is often more than one appraisal in a sentence, as demonstrated by Example 3.

Example 3 The design was deceptively simple$^{\text{Complexity}}$ and belied his ingenuity.$^{\text{Capacity}}$

An alternative strategy is to free annotators from constraints and allow multiword expressions of arbitrary length. This presents difficulties as the annotators are likely to tag units of different lengths for extremely similar expressions, but this can be compensated for by relaxing the rules for agreement by matching intersecting

[5] It would have been possible to use publicly available environments such as GATE (Cunningham et al. 2002) or the Callisto annotation tool (Day et al. 2004), but installing and customising them appropriately for this annotation task would have taken substantial effort.

annotations (Wiebe et al. 2005). Bruce and Wiebe (1999) employed yet another strategy, which was to create units from every non-compound sentence and each conjunct of every compound sentence. This is beneficial in that judges deal with precisely the same units, but it does not capture all the appraisals in expressions such as that in Example 4, in which the second conjunct contains two Appraisal-bearing expressions.

Example 4 The design was deceptively simple[Complexity] and belied his remarkable[Normality] ingenuity.[Capacity]

Our eventual chosen strategy permitted judges to annotate any number of tokens in order to allow for multiple Appraisal units of differing sizes within sentences. The judges annotated documents over two rounds, punctuated by an intermediary analysis of agreement in which they discussed examples of the most common types of disagreement, in an attempt to come to a common understanding for the second round. Annotations from the first round were left unaltered. This intermediary analysis revealed that the majority of disagreements came not from differences of opinion regarding the type of Appraisal, but rather whether an expression of Appraisal was present at all. The next section describes in detail an evaluation of inter-annotator agreement.

4 Inter-annotator agreement

As discussed in the previous section, measuring agreement is problematic as judges are liable to choose different unit lengths when marking up what is the essentially the same appraisal. Wiebe et al. (2005), who experienced this problem when annotating expressions of opinion under their own framework, accepted that it is necessary to relax matching constraints in order to consider the validity of all judges' interpretations, and therefore consider intersecting text anchors as matches. We employed the same approach in determining the inter-annotator agreement with respect to Appraisal-bearing expressions.

It is also clear that the freedom of the judges in this task requires different measures of agreement than those employed in some other types of linguistic annotation task. For example, consider how word sense annotators are obliged to choose from a limited set of senses for each token, whereas judges annotating Appraisal can potentially ascribe an extraordinarily vast range of choices. Appraisal annotators are free to select one of thirty-two classes for any contiguous substring of any length within each document; there are $16n^2 - 16n$ possible choices in a document of n tokens (approximately 6.5×10^8 possibilities in the book review corpus). This makes measuring inter-annotator agreement using conventional techniques such as Cohen's (1960) κ problematic; most of these possibilities would be left unlabelled by both annotators, counting towards agreement and diluting the effect of any disagreements.

In Wiebe et al's (2005) opinion annotation study, judges were tasked with identifying the spans of text (*text anchors*) that represent opinions. Wiebe et al.

Table 1 MUC-7 test score definitions (Chinchor 1998)

Name	Description	Calculation
COR	Number correct	
INC	Number incorrect	
MIS	Number missing	
SPU	Number spurious	
POS	Number possible	= COR + INC + MIS
ACT	Number actual	= COR + INC + SPU
F_1	F-score	= (2 × REC × PRE)/(REC + PRE)
REC	Recall	= COR/POS
PRE	Precision	= COR/ACT
SUB	Substitution	= INC/(COR + INC)
ERR	Error per response	= (INC + SPU + MIS)/(COR + INC + SPU + MIS)
UND	Under-generation	= MIS/POS
OVG	Over-generation	= SPU/ACT

measure the agreement between two judges' (a and b) sets of text anchors (A and B) as agr, a direction-sensitive measure of the proportion of A also annotated by b:

$$agr(a\|b) = \frac{|A \text{ matching B}|}{|A|} \quad (1)$$

Across the 39 documents of the Appraisal book review corpus, annotator d identified 3,176 units of appraisal, whereas j identified 2,886. Using the agr measure, d agrees with 70.6% of j's annotations while j agrees with 68.6% of d's annotations (with regard to annotated text anchors but disregarding the Appraisal type).

The 7th Message Understanding Conference (MUC-7) employed a wider range of metrics (defined in Table 1). The MUC-7 tasks included extraction of named entities, equivalence classes, attributes, facts and events (Chinchor 1998). These tasks are similar to the Appraisal-annotation task in that the units are of arbitrary length. The MUC-7 scoring system facilities the quantification of phenomena such as over-generation and under-generation, whereas agr focuses on the precision of inter-annotator agreement.

We evaluated the agreement exhibited by an annotator a as a pair-wise comparison against the other annotator b; the second annotator provides an assumed gold standard for the purposes of the agreement evaluation. Note, however, that it does not necessarily follow that REC(a w.r.t. b) = PRE(b w.r.t. a). For instance, suppose a tends to make single word annotations whilst b prefers to annotate phrases; A will contain multiple matches for some of the phrases annotated by b. The 'number correct' (COR) will differ for each annotator in the pair under evaluation.

4.1 Text anchor agreement

We first consider the level of agreement between the annotators with regard to which multiword expressions are Appraisal-bearing, regardless of their type.

Table 2 MUC-7 test scores applied to intersecting annotations

	F_1	REC	PRE	ERR	UND	OVG
d w.r.t. *j*	0.682	0.706	0.660	0.482	0.294	0.340
j w.r.t. *d*	0.715	0.667	0.770	0.444	0.333	0.230
Mean	0.698	0.686	0.711	0.462	0.312	0.274

Table 2 lists the values for the MUC-7 measures of agreement in text anchors selected by the annotators, with the harmonic mean of these scores across both annotators. The substitution rate is not listed as there is only one class when considering text anchor agreement.

The MUC-7 style measures show that annotator *d* tends to label text as Appraisal-bearing more frequently than annotator *j*. This results in higher recall for *d*, but with the usual trade-off in precision. Naturally, the opposite observation can be made about annotator *j*. Both annotators exhibit a high error rate, at 48.2 and 44.4% for *d* and *j* respectively.

4.2 Appraisal type agreement

Having examined the annotators' agreement with respect to Appraisal-bearing text anchors, we move on to analyse the agreement with respect to the Appraisal types assigned to those anchors. As above, we relaxed constraints so that annotations that overlapped were considered as matching, but with the additional constraint that the appraisal type should also match.

The Appraisal taxonomy is a hierarchical system; it is a tree with terminals corresponding to the annotation types chosen by the human judges. When investigating agreement in appraisal type we examined agreement at each level of the hierarchy; the following evaluations include not just the leaf nodes but also their parent types, collapsing the nodes into increasingly abstract semantic representations. The constituent types[6] of each of the six levels are depicted in Fig. 5.

Table 3 lists the harmonic mean MUC-7 scores at each level of the Appraisal hierarchy. As might be expected, the agreement decreases as the classes become more concrete; classes become more specific and more numerous so the complexity of the task increases. Note that there is only a small drop in agreement between levels 4 and 5 as this introduces only four new classes, and instances of these classes are infrequent. The overgeneration and undergeneration rates are not listed, as they remain constant at each level of the Appraisal hierarchy (the values of MIS, SPU, POS and ACT do not change) and so the values listed in Table 2 are correct for all levels.

The low substitution score at level 1 indicates that the annotators were able to discriminate between the three subsystems of ATTITUDE, ENGAGEMENT and

[6] Note that in these evaluations leaf nodes are included in subsequent levels. For example, FOCUS is a leaf node at level 2, but is also a member of levels 3, 4 and 5. In Fig. 5, leaf nodes are omitted from subsequent levels due to space constraints.

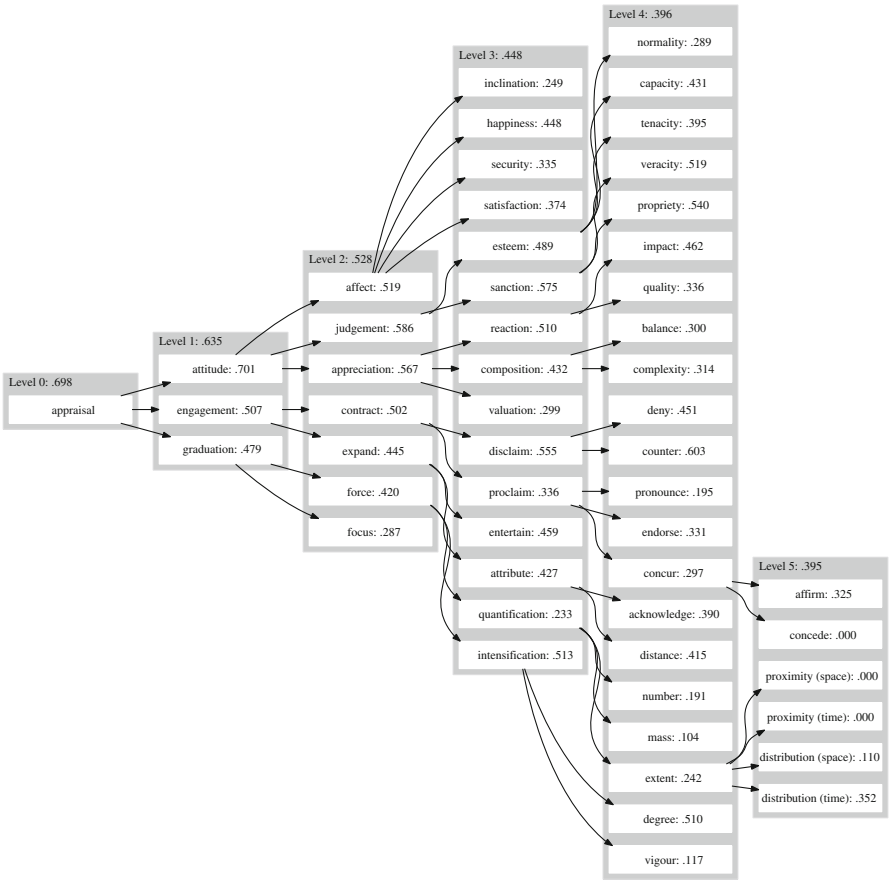

Fig. 5 The Appraisal framework showing the hierarchical levels. Labels are accompanied by the harmonic mean of the F_1 of the annotators for each appraisal type and over all types for that level

Table 3 Harmonic means of MUC-7 test scores applied to intersecting annotations and type agreement at each level of the appraisal hierarchy

Level	F_1	REC	PRE	SUB	ERR
0	0.698	0.686	0.711	0.000	0.462
1	0.635	0.624	0.647	0.090	0.511
2	0.528	0.518	0.538	0.244	0.594
3	0.448	0.441	0.457	0.357	0.655
4	0.396	0.388	0.403	0.433	0.696
5	0.395	0.388	0.403	0.433	0.696

GRADUATION; approximately 9% of annotations did not match with respect to these types. However, as the number of classes increases annotaters are more likely to disagree (as shown by the substitution rate of 43% at level 5). Similarily, the error

rate shows that the annotators frequently do not annotate intersecting spans of text (an error rate of 70% at level 5).

4.3 Measuring inter-annotator agreement beyond chance

The precision and recall scores reported in Table 2 indicate that many of the annotators' text anchors intersect. In fact, the annotators agreed that 2,223 spans of text bore some kind of appraisal. For these units, we statistically assess the inter-annotator reliability of Appraisal Type selection, using Cohen's (1960) Kappa (κ). Kappa is often employed as a measure of the agreement between pairs of annotators (Carletta 1996). It is defined as:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

where $\bar{P}$ is the proportion of agreements observed and $\bar{P}_e$ is the proportion of agreements one would expect to occur purely by chance. The denominator is thus the degree of agreement expected by chance and the numerator is the degree of agreement achieved beyond chance. κ then is 1 when there is complete agreement, 0 when there is only chance agreement and negative when there is greater disagreement than one would expect by chance.

In this scenario the proportion of agreements expected by chance, $\bar{P}_e$, is estimated using observations of the annotators' choice distribution. For a pair of annotators a and b over n annotations in $\mathbf{C}$ classes the expected proportion of agreements is:

$$\bar{P}_e = \sum_{c \in \mathbf{C}} \left(\frac{n_{a,c}}{n_a} \times \frac{n_{b,c}}{n_b} \right) \tag{3}$$

The κ values calculated at the different levels of abstraction of the Appraisal hierarchy are listed in Table 4. The values represent the reliability of agreement over all types of annotations, Attitude only, Engagement only, and Graduation only. Throughout the hierarchy the κ values indicate at least moderate agreement. As one would expect, there are better levels of agreement for types that are more abstract. The Engagement annotations exhibit reliable agreement even as the classes become increasingly concrete. When one considers the low F_1 for engagement

Table 4 κ values at the different levels of the Appraisal taxonomy over all annotation types and over Attitude, Engagement, and Graduation types only

Level	Overall	Attitude only	Engagement only	Graduation only
0	1.000	–	–	–
1	0.759	1.000	1.000	1.000
2	0.660	0.683	0.809	0.474
3	0.578	0.579	0.830	0.580
4	0.521	0.518	0.776	0.540
5	0.520	0.518	0.773	0.539

(0.507, reported in Fig. 5) it appears that the annotators have difficulty in agreeing on the presence of an Engagement annotation, but if they do so then they are able to assign the engagement type fairly reliably.

4.4 Systematic disagreement

As might be expected, some types of Appraisal are more difficult to identify than others; this is summarised by the harmonic mean of the annotators' F_1, for each class. Instances of ATTITUDE tend to be easier to identify than those of ENGAGEMENT or GRADUATION, which are similarly difficult.

The annotators did not agree on any occurrences of PROXIMITY (SPACE) or PROXIMITY (TIME) whatsoever. From the instances they marked up independently it appears the annotators held different interpretations of the concept of proximity. For example, one judge selected words that modified the distance expressed by a locution (e.g. *near, far*). In contrast the other annotator chose expressions of concepts related to proximity (e.g. *homegrown, local*). The annotators also exhibited no agreement with respect to the CONCEDE type of engagement. However, in this case the low score is perhaps due to the apparent infrequency of the class (it was annotated only once by *j* and six times by *d*).

The scores also indicate that agreement is difficult with respect to the PRONOUNCE type of engagement. In this case, the judges both selected expressions that indicate authors' conviction in a proposition (e.g. *in fact* or *it has to be said*). Judge *d*, though, saw pronouncement as being invoked whenever authors made an assertion (e.g. *this is* or *there will be*), while the selections of pronouncement made by *j* carried a strong degree of emphasis (e.g. *certainly*). There was also low agreement with respect to instances of MASS. *d* selected only strong expressions of Mass (such as *massive* or *scant*), whereas *j* also selected weaker instances such as *largely* or *slightly*. The disagreements observed in both instances of the PRONOUNCE class and instances of the MASS class are characteristic of the low agreement among many of the Appraisal classes. The judges do not tend to have extremely different interpretations of the system, but instead tend to disagree on the boundaries of a class; often, one annotator requires a greater strength of function of a word for it to be included in a class.

Another method for investigating cases of systematic disagreement between a pair of annotators is manual analysis of a contingency table. However, this can be problematic when investigating a task involving many classes. There are 32 types of Appraisal involved in this study and, when also considering instances where one annotator has not selected an intersecting string for the other's annotation, there are 1,056 contingency-pairs requiring analysis.[7]

One approach to examining systematic disagreement would be to select frequently occurring contingency pairs for further investigation, however these will be dependent on the distributions of types selected by both annotators. It is more useful to find contingency-pairs that are unexpected (those that occur differently than one would expect purely by chance). We compute the unexpectedness (u) of a

[7] The contingency tables are therefore not included here due to space constraints.

contingency-pair ($|x \in L, y \in L|$) as the difference between the observed probability and an expected probability computed from class frequency in individual annotator distributions, as represented by a matrix of all contingency-pair frequencies ($\mathbf{F}_{L,L}$), where L is the set of labels in the annotation problem:

$$u(|x \in L, y \in L|) = P(O) - P(E) \tag{4}$$

$$P(O) = \frac{\mathbf{F}_{x,y}}{\sum_{l \in L} \sum_{m \in L} \mathbf{F}_{l,m}} \tag{5}$$

$$P(E) = \frac{\sum_{l \in L} \mathbf{F}_{x,l}}{\sum_{l \in L} \sum_{m \in L} \mathbf{F}_{l,m}} \times \frac{\sum_{m \in L} \mathbf{F}_{m,y}}{\sum_{l \in L} \sum_{m \in L} \mathbf{F}_{l,m}} \tag{6}$$

The resulting unexpectedness value is greater than zero if a contingency pair occurs more than one would expect by chance, zero if it occurs as one would expect by chance, and less than zero if it occurs less frequently than expected by chance. To reduce the scale of a manual search we investigated contingency-pairs where the unexpectedness value was greater than the mean plus the standard deviation of all unexpectedness values ($u > \bar{x} + \sigma$). This is an arbitrary selection but seems to include interesting contingency-pairs while keeping the manual search manageable.

For instance, of the instances of QUALITY selected by j, d chooses IMPACT for 19% ($u = 0.181$) and VALUATION for 22% ($u = 0.191$). As these three classes are very closely related in the Appraisal taxonomy, it is unsurprising that the annotators should disagree about their instances. Similarly, the closely related pairs of CAPACITY and TENACITY, (e.g. *single-minded thoroughness*), COMPLEXITY and BALANCE (e.g. *dichotomies of character*), and DEGREE and FOCUS (e.g. *authentically*) were also difficult to discriminate.

Other apparent examples of disagreement arise from the problems caused by the flexibility of the coding scheme. For example, 33% of d's annotations of PROXIMITY (SPACE) were ascribed to CAPACITY by j ($u = 0.250$). The high percentage is due to the rarity of annotations of PROXIMITY (SPACE), while the disagreement itself comes from the annotators selecting units of differing length, as shown in Examples 5 and 6.

Example 5 [d] But at key points in this story, one gets the feeling that the essential factors are operating just outside$^{\text{PROXIMITY(SPACE)}}$ James's field of vision.$^{\text{CAPACITY}}$

Example 6 But at key points in this story, one gets the feeling that the essential factors are operating just outside James's field of vision.$^{\text{CAPACITY}}$

Another interesting case of frequent disagreement is the pair of SATISFACTION and PROPRIETY. Even though they are not closely related in the Attitude subsystem, j chooses PROPRIETY for 21% of d's annotations of SATISFACTION ($u = 0.188$). Examples 4 and 4 typify this confusion, in which there is disagreement with respect to the subject of the appraisal rather than the type of appraisal. Annotator d's selection is intepreting the sentence as the artist experiencing negative SATISFACTION in response to the critics, whereas j's selection interprets it as the author reproaching the critics for their treatment of the artist.

Example 7 [d] Like him, Vermeer—or so he chose to believe—was an artist neglected[SATISFACTION] wronged[SATISFACTION] by critics and who had died an almost unknown.

Example 8 [j] Like him, Vermeer—or so he chose to believe—was an artist neglected and wronged[PROPRIETY] by critics and who had died an almost unknown.

These examples illustrate a shortcoming of the coding scheme which assumes that only one type of Appraisal is conveyed by each appraisal-bearing unit.

5 Ambiguous Appraisal-bearing expressions

Disagreement between the two coders concerning certain words was apparent throughout the data annotation study. During the intermediate analysis period the judges discussed several examples of disagreement that they could not resolve; both annotators were able to understand the other's point of view and so the instance remained ambiguous. In many cases, this was due to different interpretations of the object under appraisal. For example, consider:

Example 9 Knutie learns to forgive Max, Summer Feelin's father, for abandoning her.

Annotator *d* read this example as expressing Knutie's emotional reaction at being abandoned and annotated the phrase with SATISFACTION. Annotator *j* instead interpreted the sentence as a judgement of Max's character and labelled it with PROPRIETY.

Both interpretations seem reasonable, and so to investigate this further we produced a questionnaire to determine whether it is possible for judges to reach a consensus on the Appraisal types of ambiguous terms in specific contexts. The questionnaire presented thirty sentences containing a form of the word *abandon* and asked respondents to choose one of six options for each sentence:

1. an emotion;
2. a judgement about the reliability of a person (or group of people);
3. a judgement about the morality of a person (or group of people);
4. many of the above;
5. none of the above; or
6. unsure.

Example sentences were selected at random from the British National Corpus (Leech 1992), but following the part-of-speech distribution of the *abandon* inflectional and derivational variants observed in the BNC. The categories were selected based on intuitions as to which of the thirty-two classes of appraisal were likely to be chosen by the respondents. Respondents were also asked if they spoke English as their first language, and if they were familiar with the linguistic theory of Appraisal. The questionnaire was advertised through word-of-mouth and via the Appraisal Analysis discussion group.[8] Forty-seven respondents completed the survey.

[8] http://tech.groups.yahoo.com/group/AppraisalAnalysis/.

5.1 Measuring agreement amongst many annotators

In order to measure agreement between the respondents to the questionnaire we employed Fleiss's (1971) κ, a variant of Cohen's (1960) κ which applies the measurement to arbitrary but fixed numbers of annotators. Let N be the total number of sentences, let n be the number of respondents and let k be the number of categories. n_{ij} is the number of respondents who assigned the i-th sentence to the j-th category. Fleiss expanded Cohen's κ as:

$$\kappa = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn\left[1 + (n-1)\sum_{j=1}^{k} p_j^2\right]}{Nn(n-1)\left(1 - \sum_{j=1}^{k} p_j^2\right)} \tag{7}$$

Fleiss (1971) also demonstrated how to estimate the variance of κ, if one assumes that N (in this application, the number of sentences) is large enough for the proportions of assignments to each category (p_j) to be constant. Under this assumption κ is a function of only the random variable $\sum_i \sum_j n_{ij}^2$ and Fleiss (1971) calculated its variance as:

$$\mathrm{Var}(\kappa) = \frac{2}{Nn(n-1)} \times \frac{\sum_j p_j^2 - (2n-3)\left(\sum_j p_j^2\right)^2 + 2(n-2)\sum_j p_j^3}{\left(1 - \sum_j p_j^2\right)^2} \tag{8}$$

Comparing $\kappa / \sqrt{\mathrm{Var}(\kappa)}$ to tables of the Normal distribution provides an estimate of the significance of κ.

Fleiss (1971) also defined κ_j, a measure of the agreement exhibited by the respondents over the j-th category:

$$\kappa_j = \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j\left[1 + (n-1)p_j\right]}{Nn(n-1)p_j(1-p_j)} \tag{9}$$

The approximate variance of κ_j is given by:

$$\mathrm{Var}(\kappa_j) = \frac{\left(1 + 2(n-1)p_j\right)^2 + 2(n-1)p_j(1-p_j)}{Nn(n-1)^2 p_j(1-p_j)} \tag{10}$$

5.2 Agreement in the ambiguous term categorisation task

Table 5 lists the κ scores for each combination of the sets of respondents who speak English as a first language, those who do not, those who are familiar with Appraisal and those who are not. Clearly the task is difficult since $\kappa = 0.135$ over all categories and all respondents. Landis and Koch (1977) state that $\kappa < 0.200$ should be considered as 'poor agreement'. Nevertheless it is significantly more than chance, demonstrating that the respondents are using the context provided by the sentence to select the category they believe is appropriate. Respondents seem to find it easier to agree on sentences where there is an absence of Appraisal, as

Table 5 All possible combinations of questionnaire respondents with English as a first language and/or familiarity with Appraisal, with corresponding respondent frequencies (n) and Kappa (κ) scores

English?	Appraisal?	n	K	K_{emotion}	K_{morality}	$K_{\text{reliability}}$	K_{many}	K_{none}	K_{unsure}
*	*	47	0.135	0.094	0.084	0.079	0.034	0.259	0.047
*	Yes	26	0.143	0.074	0.104	0.072	0.029	0.265	0.113
*	No	21	0.116	0.129	0.052	0.078	0.024	0.230	−0.010
Yes	*	34	0.145	0.077	0.090	0.080	0.043	0.284	0.059
Yes	Yes	17	0.178	0.036	0.113	0.078	0.049	0.324	0.174
Yes	No	17	0.109	0.117	0.060	0.069	0.021	0.226	−0.011
No	*	13	0.112	0.168	0.062	0.045	0.007	0.190	0.012
No	Yes	9	0.087	0.142	0.061	0.021	−0.006	0.155	0.012
No	No	4	0.109	0.198	−0.049	0.019	−0.043	0.216	−0.008

All values are significant at $p < 0.05$

demonstrated by the higher values for κ_{none} [since $0.200 < \kappa < 0.400$ can be considered as 'fair agreement' (Landis and Koch 1977)].

As might be anticipated, the best performing subset of respondents was those who speak English as a first language and have a knowledge of Appraisal ($\kappa = 0.178$). The group with the least agreement was those who do not speak English as a first language but are familiar with Appraisal. This suggests that native proficiency in English is more useful than knowledge of Appraisal when completing this task.

The group of those familiar with Appraisal exhibit the lowest (significant) κ_{emotion} at 0.077. This same group rarely selects the Emotion category (4.6% of annotations). This perhaps indicates that AFFECT is not a suitable Appraisal class for the *abandon* word family. Instead, this group of Appraisal experts prefers to ascribe TENACITY (22.2%) or PROPRIETY (12.5%) to the example sentences.

This exercise demonstrates the complexity of some instances of Appraisal annotation. In particularly ambiguous cases, such as the choices posed by the word *abandon*, even experts in Appraisal find it difficult to agree on classes. For instance, the classifications for the sentence in Example 10 were particularly divergent, with 31.9% of respondents choosing TENACITY and 42.6% selecting PROPRIETY.

Example 10 To crush strikes and **abandon** political reform would be to throw himself into the arms of those groups wedded not just to authoritarian politics but to neo-Stalinist economic institutions and principles.

6 A gold-standard for Appraisal analysis

Despite the disagreements outlined above, the annotators d and j do frequently agree on instances of types in the Appraisal framework. These instances may be useful to researchers engaging in Appraisal research, so it is appropriate to collate the annotations into machine- and human- readable formats.[9] We collated a gold

[9] A script to download the articles and apply the annotations is available for download from http://folk.uio.no/jread/resources/appcor.tar.gz.

Table 6 The accuracy of support vector machine classifiers applied to labelling Appraisal-bearing expressions from the gold standard, with two baselines listed for comparison: choosing the *Majority* class in the training data, and choosing a class at *random*

Level	SVM	Majority	Random
1	0.824	0.742	0.333
2	0.441	0.409	0.143
3	0.351	0.232	0.063
4	0.326	0.115	0.036
5	0.326	0.115	0.031

The labels at each level correspond to those in Fig. 5

standard for each level of the Appraisal hierarchy by searching both annotators' selections for matching pairs. Two annotations formed a pair if their spans intersected and their labelled type matched, or shared a common ancestor in one of the Appraisal subsystems. Our analysis indicated only a small difference in rates of inter-annotator agreement between the two rounds of annotation, hence the gold-standard includes annotations from both rounds.

The gold standard consists of six XML documents, each corresponding to a level in the Appraisal hierarchy depicted in Fig. 5. Each XML document is headed by a CORPUS element, which contains several TEXT elements which contain the complete book review texts. Elements within that text indicate the annotations where appropriate; the elements are named for the Appraisal type, as appropriate to the hierarchical level represented by the particular document. ATTITUDE annotation elements contain a POLARITY attribute (POSITIVE or NEGATIVE), and GRADUATION annotation elements contain a SCALING attribute (UP or DOWN), in cases where the annotators agreed on the polarity or direction.

We conducted a preliminary experiment using the gold-standard data in order to assess the viability of automatically classifying expressions of Appraisal. Our experiment employed Thorsten Joachim's implementation[10] of multiclass support vector machines as described by Crammer and Singer (2001). Models were trained for each level of the Appraisal heirarchy using the development data and tested with the gold standard. The accuracy of each model is listed in Table 6, along with baseline accuracies obtained by choosing the *majority* class, or by choosing a class at *random*. The results indicate that computational classification of Appraisal-bearing expressions is feasible, with the support vector machines outperforming all baselines.

7 Conclusions

In this article we have reviewed Appraisal, a systemic functional linguistic theory of evaluation in text. The theory describes a typology of language, consisting of three subsystems that operate in parallel: ATTITUDE describes the language used to communicate personal feelings in terms of emotional reactions, judgements of

[10] Available from http://svmlight.joachims.org/svm_multiclass.html.

people and appreciation of objects; ENGAGEMENT considers the positioning of oneself with respect to the opinions of others; and GRADUATION is concerned with how language can function to amplify or diminish the attitude or engagement conveyed by text. We proposed applying Appraisal theory to tasks in sentiment analysis because of the theory's detailed consideration not only of types of evaluation and modifiers of the strength of evaluation, but also of how writers report the opinions of other people.

In order to create gold standard data for Appraisal analysis methods, we conducted an Appraisal annotation exercise. The corpus used consisted of thirty-eight book reviews, as articles from this domain are likely to contain examples of each of Appraisal's many classes. Two human judges carried out the annotation task, annotating text independently of one another. They were instructed to select Appraisal-bearing terms, and label them with one of the 32 types of Appraisal. They were also asked to assign a polarity (positive or negative) to ATTITUDE-bearing expressions and a scaling (up or down) to GRADUATION-bearing expressions.

Judges were permitted to annotate any number of contiguous tokens. This meant it was not immediately possible to measure inter-annotator agreement using standard techniques such as Cohen's (1960) κ, as there are so many possible labellings in the corpus. The vast majority of possibilities would be instances left unlabelled by both annotators. These would be included in the κ measurement of agreement, and so would wash out the effects of any disagreements. Instead the agreement study employed metrics taken from the 7th Message Understanding Conference (MUC-7). The tasks in MUC-7 are similar to the Appraisal annotation task in that expressions are of an arbitrary number of tokens, and so suited to measuring annotator agreement as a pair-wise comparison, taking one annotator as the 'system' and the other as the 'standard'. The MUC-7 metrics were used to measure the agreement both in strings of words annotated by the annotators and in Appraisal type. Agreement was analysed at each level of the Appraisal hierarchy, where concrete terms were collapsed into their parents nodes for increasingly abstract types. The agreement exhibited between the annotators dropped as the classes became more concrete. Naturally, some classes were easier to agree upon than others; types of ATTITUDE were easiest to identify.

Taking the intersection of both annotators' sets of annotated strings enabled analysis of agreement beyond chance using Cohen's κ. This revealed that there was at least moderate agreement beyond chance at all levels of the hierarchy. In general there was more agreement for types that were more abstract. This did not hold for types of Engagement, which exhibited strong agreement even at the most concrete levels of the hierarchy.

Some instances of disagreement indicated shortcomings in the annotation framework, in that only one appraisal type was permitted for each expression. However, in several instances both annotators made reasonable but conflicting decisions for the same expression. These were due to a number of reasons which need to be considered in future Appraisal annotation studies: words may be relevant to multiple classes; the class of a word can vary according to its context; appraisals may relate to multiple entities (c.f. Examples 7 and 8); and interpretation of Appraisal-bearing words is subjective.

Despite the disagreements however, the annotators did frequently agree on annotations of the Appraisal framework. We created a gold-standard for each level of the Appraisal framework by selecting instances of agreement from the annotators' sets. We also created a development data set from the symmetric difference of the annotators' sets; while this data is not reliable it still contains useful information about Appraisal. Training SVM classifiers on this development data resulted in models that performed better than naïve baselines.

In on-going work we are continuing to employ this corpus as a gold-standard for Appraisal classification experiments. In particular, we are adapting methods such as Turney's (2002) SO-PMI-IR to classify words according to the classes of the Appraisal hierarchy. Developments of the gold-standard will expand the annotation scheme by considering appraisals with multiple subjects.

The automatic identification of Appraisal-bearing expressions is the first step in developing Appraisal-aware approaches to sentiment analysis. Such approaches may employ heuristics based on the Appraisal types, much like the contextual valence shifters described by Polanyi and Zaenen (2006) and the models for the compositionality of sentiment proposed by Moilanen and Pulman (2007).

The basis of such heuristics would be the positivity or negativity of instances of ATTITUDE. Heuristics describing the effects of GRADUATION are fairly straightforward to derive, with up-scaling items intensifying the polarity and down-scaling items diminishing the polarity. The effects implied by different types of ENGAGEMENT are more complex, however. For example, intuitively one might presume instance of DENY would nullify any associated polarity, whereas ENDORSE would perform intensification. These relationships are not formally specified by Appraisal theory, so an analysis of the correlations between expression-level polarity, instances of ENGAGEMENT and sentence-level polarity in the corpus described in this paper is an interesting area for future work.

Acknowledgments We would like to express our gratitude to David Hope, who kindly served as one of the annotators in this study. Our thanks also go to Bill Keller, who provided advice with respect to the design of our annotation scheme, and to the reviewers for their advice on improving this article. The work of the first author was supported by a studentship provided by the Engineering and Physical Sciences Research Council (United Kingdom) and was conducted at the University of Sussex.

References

Argamon, S., Bloom, K., Esuil, A., & Sebastiani, F. (2007). Automatically determining attitude type and force for sentiment analysis. In *Proceedings of the 3rd language and technology conference (LTC'07), Poznan, PL* (pp. 369–373).

Bruce, R., & Wiebe, J. (1999). Recognizing subjectivity: A case study in manual tagging. *Natural Language Engineering, 5*(1), 1–16.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Chinchor, N. (1998). MUC-7 test scores introduction. In *Proceedings of the seventh message understanding conference*

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measures, 20*, 37–46.

Crammer, K., & Singer, Y. (2001). On the algorithmic implementation of multi-class svms. *Journal of Machine Learning Research, 2*, 265–292.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*.

Day, D., McHenry, C., Kozierok, R., & Riek, L. (2004). Callisto: A configurable annotation workbench. In *LREC 2004: Fourth international conference on language resources and evaluation, Lisbon, Portugal*.

Di Marco, C., & Mercer, R.E. (2004). Using hedges to classify citations in scientific articles. In *Computing attitude and affect in text: Theory and applications, Springer, Netherlands* (pp. 247–263).

Ekman, P. (1993). Facial expression of emotion. *American Psychologist, 48*, 384–392.

Esuli, A., & Sebastiani, F. (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of the 14th ACM international conference on information and knowledge management (CIKM'05), Bremen, DE*.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Gratch, J., & Marsella, S. (2004). A domain-independent framework for modeling emotion. *Cognitive Systems Research, 5*(4), 269–306.

Hyland K. (1998). Hedging in scientific research articles. Amsterdam/Philadelphia: John Benjamins.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. Longman: London.

Labov, W. (1984). Intensity. In: D. Schiffrin (Ed.), *Meaning, form, and use in context: Linguistic applications*. Washington, D.C.: Georgetown University Press.

Landis, J. R., & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Leech, G. (1992). 100 million words of English: The British National Corpus. *Language Research, 28*, 1–13.

Martin, J. R. (2004). Mourning: How we get aligned. *Discourse & Society, 15*(2–3), 321–344.

Martin, J. R., & White, P. R. R. (2005). Language of evaluation: Appraisal in English. London: Palgrave Macmillan.

Mihalcea, R., & Chklovski, T. (2003). Open mind word expert: Creating large annotated data collections with web users' help. In *Proceedings of the EACL 2003 workshop on linguistically annotated corpora (LINC 2003), Budapest, Hungary*.

Moilanen, K., & Pulman, S. (2007). Sentiment composition. In *Proceedings of recent advances in natural language processing (RANLP 2007), Borovets, Bulgaria* (pp. 378–382).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 conference on empirical methods in natural language processing, Philadelphia, PA, USA*.

Polanyi, L., & Zaenen, A. (2006). Contextual valence shifters. In *Computing attitude and affect in text: Theory and applications* (pp. 1–10). Springer.

Read, J. (2004). *Recognising affect in text using pointwise mutual information*. Master's thesis, University of Sussex.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop, Ann Arbor, MI, USA*.

Scherer, K. R., Schoor, A., & Johnstone, T. (Eds.) (2001). *Appraisal processes in emotion: Theory, methods, research*. Canary, NC: Oxford University Press.

Strapparava, C., & Mihalcea, R. (2007). Sem Eval-2007 task 14: Affective text. In *Proceedings of the 4th international workshop on semantic evaluations (SemEval 2007), Prague, Czech Republic*.

Subasic, P., & Huettner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems, 9*(4), 483–496.

Taboada, M., & Grieve, J. (2004). Analyzing appraisal automatically. In *Spring symposium on exploring attitude and affect in text, American Association for Artificial Intelligence, Stanford, aAAI Technical Report SS-04-07*.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA*.

White, P. R. R. (2002). Appraisal—The language of evaluation and stance. In J. Verschueren, J. O. Östman, J. Blommaert & C. Bulcaen (Eds.), *Handbook of pragmatics* (pp. 1–27). Amsterdam: John Benjamins.

Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. In *Proceedings of the conference on information and knowledge management (CIKM)*.

Wiebe, J., Breck, E., Buckley, C., Cardie, C., Davis, P., Fraser, B., et al. (2003). Recognizing and organizing opinions expressed in the world press. In *Proceedings of the AAAI spring symposium on new directions in question answering*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277–308.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2–3), 165–210.

Yang, C., Lin, K. H. Y., & Chen, H. H. (2007). Building emotion lexicon from weblog corpora. In *Proceedings of the ACL 2007 demo and poster sessions, Prague, Czech Republic* (pp. 133–136).