Annotating Expressions of Appraisal in English

Jonathon Read, David Hope and John Carroll
Department of Informatics
University of Sussex
United Kingdom
{j.l.read,drh21,j.a.carroll}@sussex.ac.uk

Abstract

The Appraisal framework is a theory of the language of evaluation, developed within the tradition of systemic functional linguistics. The framework describes a taxonomy of the types of language used to convey evaluation and position oneself with respect to the evaluations of other people. Accurate automatic recognition of these types of language can inform an analysis of document sentiment. This paper describes the preparation of test data for algorithms for automatic Appraisal analysis. The difficulty of the task is assessed by way of an inter-annotator agreement study, based on measures analogous to those used in the MUC-7 evaluation.

1 Introduction

The Appraisal framework (Martin and White, 2005) describes a taxonomy of the language employed in communicating evaluation, explaining how users of English convey attitude (emotion, judgement of people and appreciation of objects), engagement (assessment of the evaluations of other people) and how writers may modify the strength of their attitude/engagement. Accurate automatic analysis of these aspects of language will augment existing research in the fields of sentiment (Pang et al., 2002) and subjectivity analysis (Wiebe et al., 2004), but assessing the usefulness of analysis algorithms leveraging the Appraisal framework will require test data.

At present there are no machine-readable Appraisal-annotated texts publicly available. Real-world instances of Appraisal in use are limited to example extracts that demonstrate the theory, coming from a wide variety of genres as disparate as news reporting (White, 2002; Martin, 2004) and poetry (Martin and White, 2005). These examples, while useful in demonstrating the various aspects of Appraisal, can only be employed in a qualitative analysis and would bring about inconsistencies if analysed collectively — one can expect the writing style to depend upon the genre, resulting in significantly different syntactic constructions and lexical choices.

We therefore need to examine Appraisal across documents in the same genre and investigate patterns within that particular register. This paper discusses the methodology of an Appraisal annotation study and an analysis of the inter-annotator agreement exhibited by two human judges. The output of this study has the additional benefit of bringing a set of machine-readable annotations of Appraisal into the public domain for further research.

This paper is structured as follows. The next section offers an overview of the Appraisal framework. Section 3 discusses the methodology adopted for the annotation study. Section 4 discusses the measures employed to assess inter-annotator agreement and reports the results of these measures. Section 5 offers an analysis of cases of systematic disagreement. Other computational work utilising the Appraisal framework is reviewed in Section 6. Section 7 summarises the paper and outlines future work.

2 The linguistic framework of Appraisal

The Appraisal framework (Martin and White, 2005) is a development of work in Systemic Functional

Proceedings of the Linguistic Annotation Workshop, pages 93–100,
Prague, June 2007.

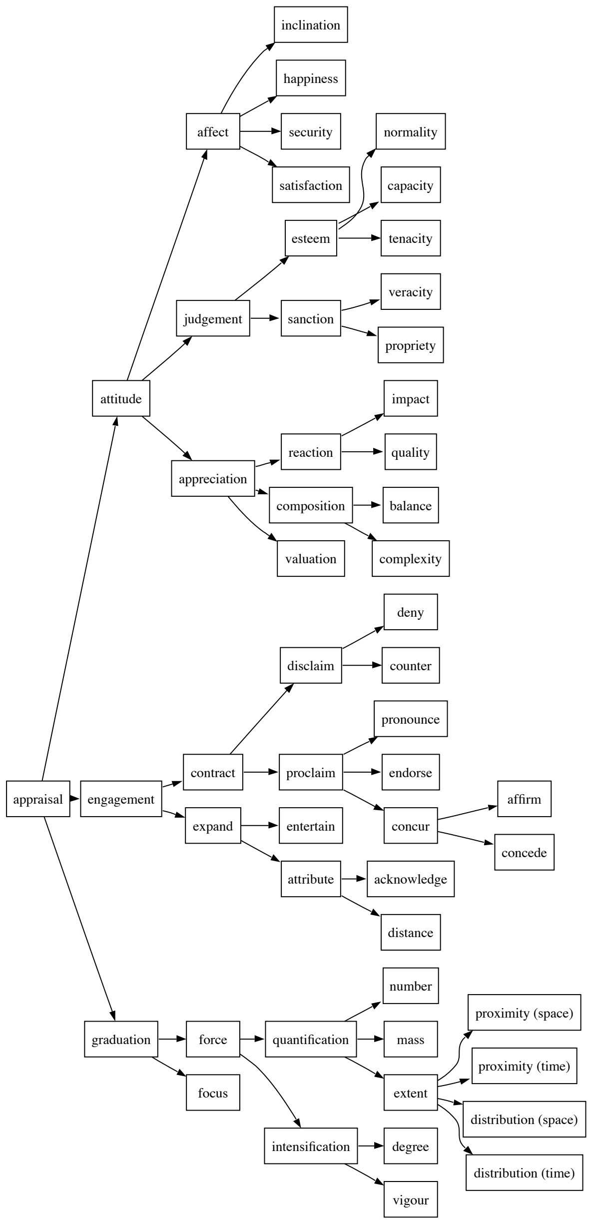

Figure 1: The Appraisal framework.

Linguistics (Halliday, 1994) and is concerned with interpersonal meaning in text—the negotiation of social relationships by communicating emotion, judgement and appreciation. The taxonomy described by the Appraisal framework is depicted in Figure 1.

Appraisal consists of three subsystems that operate in parallel: **attitude** looks at how one expresses private state (Quirk et al., 1985) (one's emotion and opinions); **engagement** considers the positioning of oneself with respect to the opinions of others and **graduation** investigates how the use of language functions to amplify or diminish the attitude and engagement conveyed by a text.

2.1 Attitude: emotion, ethics and aesthetics

The Attitude sub-system describes three areas of private state: emotion, ethics and aesthetics. An attitude is further qualified by its polarity (*positive* or *negative*). Affect identifies feelings—author's emotions as represented by their text. Judgement deals with authors' attitude towards the behaviour of people; how authors applaud or reproach the actions of others. Appreciation considers the evaluation of things—both man-made and natural phenomena.

2.2 Engagement: appraisals of appraisals

Through engagement, Martin and White (2005) deal with the linguistic constructions by which authors construe their point of view and the resources used to adopt stances towards the opinions of other people. The theory of engagement follows Stubbs (1996) in that it assumes that all utterances convey point of view and Bakhtin (1981) in supposing that all utterances occur in a miscellany of other utterances on the same motif, and that they carry both implicit and explicit responses to one another. In other words, all text is inherently dialogistic as it encodes authors' reactions to their experiences (including previous interaction with other writers). Engagement can be both retrospective (that is, an author will acknowledge and agree or disagree with the stances of others who have previously appraised a subject), and prospective (one may anticipate the responses of an intended audience and include counter-responses in the original text).

2.3 Graduation: strength of evaluations

Martin and White (2005) consider the resources by which writers alter the strength of their evaluation as a system of graduation. Graduation is a general property of both attitude and engagement. In attitude it enables authors to convey greater or lesser degrees of positivity or negativity, while graduation of engagements scales authors' conviction in their utterance.

Graduation is divided into two subsystems. Force alters appraisal propositions in terms of its intensity, quantity or temporality, or by means of spatial metaphor. Focus considers the resolution of semantic categories, for example:

They play *real* jazz.
They play jazz, *sort of*.

In real terms a musician either plays jazz or they do not, but these examples demonstrate how authors blur the lines of semantic sets and how binary relationships can be turned into scalar ones.

3 Annotation methodology

The corpus used in this study consists of unedited book reviews. Book reviews are good candidates for this study as, while they are likely to contain similar language by virtue of being from the same genre of writing, we can also expect examples of Appraisal's many classes (for example, the emotion attributed to the characters in reviews of novels, judgements of authors' competence and character, appreciation of the qualities of books and engagement with the propositions put forth by the authors under review).

The articles were taken from the web sites of four British newspapers (The Guardian, The Independent, The Telegraph and The Times) on two different dates—31 July 2006 and 11 September 2006. Each review is attributed to a unique author. The corpus is comprised of 38 documents, containing a total of 36,997 tokens in 1,245 sentences.

Two human annotators, d and j, participated in this study, assigning tags independently. The annotators were well-versed in the Appraisal framework, having studied the latest literature. The judges were asked to annotate appraisal-bearing terms with the appraisal type presumed to be intended by the author of the text. They were asked to highlight each example of appraisal and specify the type of attitude, engagement or graduation present. They also assigned a *polarity* (positive or negative) to attitudinal items and a *scaling* (up or down) to graduating items, employing a custom-developed software tool to annotate the documents.

Four alternative annotation strategies were considered. One approach is to allow only a single token per annotation. However, this is too simplistic for an Appraisal annotation study—a unit of Appraisal is frequently larger than a single token. Consider the following examples:

(1)
The design was *deceptively*–VERACITY *simple*–COMPLEXITY. (∗)

(2)
The design was *deceptively simple*–COMPLEXITY.

Example 1 demonstrates that a single-token approach is inappropriate as it ascribes a judgement of someone's honesty, whereas Example 2 indicates the correct analysis—the sentence is an appreciation of the simplicity of the "design". This example shows how it is necessary to annotate larger units of appraisal-bearing language.

Including more tokens, 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 tagging entire sentences in order to constrain the annotators' range of choices. This resulted in its own problems as there is often more than one appraisal in a sentence, for example:

(3)
The design was *deceptively simple*–COMPLEXITY and belied his *ingenuity*–CAPACITY.

An alternative approach is to permit annotators to tag an arbitrary number of contiguous tokens. Arbitrary-length tagging is disadvantageous as the judges will frequently tag units of differing length, but this can be compensated for by relaxing the rules for agreement—for example, by allowing intersecting annotations to match successfully (Wiebe et al., 2005). Bruce and Wiebe (1999) employ another approach, creating units from every non-compound sentence and each conjunct of every compound sentence. This side-steps the problem of ambiguity in appraisal unit length, but will still fail to capture both appraisals demonstrated in the second conjunct of Example 4.

(4)
The design was *deceptively simple*–COMPLEXITY and belied his *remarkable*–NORMALITY *ingenuity*–CAPACITY.

Ultimately in this study, we permitted judges to annotate any number of tokens in order to allow for multiple Appraisal units of differing sizes within sentences. Annotation was carried out over two rounds, punctuated by an intermediary analysis of

	d	j		d	j		d	j
Inclination	1.26	3.50	Balance	2.64	1.84	Distance	0.69	0.59
Happiness	2.80	2.32	Complexity	2.52	2.74	Number	0.82	2.63
Security	4.31	2.22	Valuation	6.08	9.29	Mass	0.22	1.63
Satisfaction	1.67	2.32	Deny	3.05	3.67	Proximity (Space)	0.09	0.14
Normality	8.00	4.44	Counter	4.79	3.78	Proximity (Time)	0.03	0.55
Capacity	11.46	9.63	Pronounce	3.84	1.21	Distribution (Space)	0.41	1.39
Tenacity	3.72	4.44	Endorse	2.05	1.49	Distribution (Time)	0.82	2.56
Veracity	3.15	2.01	Affirm	0.54	1.14	Degree	4.38	5.72
Propriety	13.32	12.61	Concede	0.38	0.03	Vigour	0.60	0.45
Impact	6.11	4.23	Entertain	2.27	2.43	Focus	3.02	2.29
Quality	2.55	3.40	Acknowledge	2.42	3.33			

Table 1: The distribution of the Appraisal types selected by each annotator (%).

	d	j
Documents	115.74	77.21
Sentences	3.65	2.43
Words	0.12	0.08

Table 2: The density of annotations relative to the number of documents, sentences and words.

agreement and disagreement between the two annotators. The judges discussed examples of the most common types of disagreement in an attempt to acquire a common understanding for the second round, but annotations from the first round were left unaltered.

Following the methodology described above, d made 3,176 annotations whilst j made 2,886 annotations. The distribution of the Appraisal types ascribed is shown in Table 1, while Table 2 details the density of annotations in documents, sentences and words.

4 Measuring inter-annotator agreement

The study of inter-annotator agreement begins by considering the level of agreement exhibited by the annotators in deciding which tokens are representative of Appraisal, irrespective of the type. As discussed, this is problematic as judges are liable to choose different length token spans when marking up what is essentially the same appraisal, as demonstrated by Example 5.

(5)
[d] It is tempting to point to the bombs in London and elsewhere, to the *hideous mess*–QUALITY in Iraq, to recent victories of the Islamists, to the *violent and polarised rhetoric*–PROPRIETY and answer yes.

[j] It is tempting to point to the bombs in London and elsewhere, to the *hideous*–QUALITY *mess*–BALANCE in Iraq, to recent victories of Islamists, to the *violent*–PROPRIETY and *polarised*–PROPRIETY rhetoric and answer yes.

Wiebe et al. (2005), who faced this problem when annotating expressions of opinion under their own framework, accept that it is necessary to consider the validity of all judges' interpretations and therefore consider intersecting annotations (such as "hideous" and "hideous mess") to be matches. The same relaxation of constraints is employed in this study.

Tasks with a known number of annotative units can be analysed with measures of agreement such as Cohen's κ Coefficient (1960), but the judges' freedom in this task prohibits meaningful application of this measure. For example, consider how word sense annotators are obliged to choose from a limited fixed set of senses for each token, whereas judges annotating Appraisal are free to select one of thirty-two classes for any contiguous substring of any length within each document; there are $16\left(n^2 - n\right)$ possible choices in a document of n tokens (approximately 6.5×10^8 possibilities in this corpus).

A wide range of evaluation metrics have been employed by the Message Understanding Conferences (MUCs). The MUC-7 tasks included extraction of named entities, equivalence classes, attributes, facts and events (Chinchor, 1998). The participating systems were evaluated using a variety of related measures, defined in Table 3. These tasks are similar to Appraisal annotation in that the units are formed of an arbitrary number of contiguous tokens.

In this study the agreement exhibited by an annotator a is evaluated as a pair-wise comparison against the other annotator b. Annotator b provides

COR	Number correct	
INC	Number incorrect	
MIS	Number missing	
SPU	Number spurious	
POS	Number possible	$= \text{COR} + \text{INC} + \text{MIS}$
ACT	Number actual	$= \text{COR} + \text{INC} + \text{SPU}$
FSC	F-score	$= (2 \times \text{REC} \times \text{PRE}) / (\text{REC} + \text{PRE})$
REC	Precision	$= \text{COR}/\text{POS}$
PRE	Recall	$= \text{COR}/\text{ACT}$
SUB	Substitution	$= \text{INC}/(\text{COR} + \text{INC})$
ERR	Error per response	$= (\text{INC} + \text{SPU} + \text{MIS}) / (\text{COR} + \text{INC} + \text{SPU} + \text{MIS})$
UND	Under-generation	$= \text{MIS}/\text{POS}$
OVG	Over-generation	$= \text{SPU}/\text{ACT}$

Table 3: MUC-7 score definitions (Chinchor 1998).

	FSC	**REC**	**PRE**	**ERR**	**UND**	**OVG**
d	0.682	0.706	0.660	0.482	0.294	0.340
j	0.715	0.667	0.770	0.444	0.333	0.230
$\bar{x}$	0.698	0.686	0.711	0.462	0.312	0.274

Table 4: MUC-7 test scores, evaluating the agreement in text anchors selected by the annotators. $\bar{x}$ denotes the average value, calculated using the harmonic mean.

a presumed gold standard for the purposes of evaluating agreement. Note, however, that in this case it does not necessarily follow that REC (a w.r.t. b) $=$ PRE (b w.r.t. a). Consider that a may tend to make one-word annotations whilst b prefers to annotate phrases; the set of a's annotations will contain multiple matches for some of the phrases annotated by b (refer to Example 5, for instance). The 'number correct' will differ for each annotator in the pair under evaluation.

Table 4 lists the values for the MUC-7 measures applied to the text spans selected by the annotators. Annotator d is inclined to identify text as Appraisal more frequently than annotator j. This results in higher recall for d, but with lower 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. The substitution rate is not listed as there are no classes to substitute when considering only text anchor agreement. The second round of annotation achieved slightly higher agreement (the mean F-score increased by 0.033).

	FSC	**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

Table 5: Harmonic means of the MUC-7 test scores evaluating the agreement in text anchors and Appraisal classes selected by the annotators, at each level of hierarchical abstraction.

Having considered the annotators' agreement with respect to text anchors, we go on to analyse the agreement exhibited by the annotators with respect to the types of Appraisal assigned to the text anchors. The Appraisal framework is a hierarchical system—a tree with leaves corresponding to the annotation types chosen by the judges. When investigating agreement in Appraisal type, the following measures include not just the leaf nodes but also their parent types, collapsing the nodes into increasingly abstract representations. For example *happiness* is a kind of *affect*, which is a kind of *attitude*, which is a kind of *appraisal*. These relationships are depicted in full in Figure 2. Note that in the following measurements of inter-annotator agreement leaf nodes are included in subsequent levels (for example, *focus* is a leaf node at level 2, but is also considered to be a member of levels 3, 4 and 5).

Table 5 shows the harmonic means of the MUC-7 measures of the annotators' agreement at each of the levels depicted in Figure 2. As one might expect, the agreement steadily drops as the classes become more concrete—classes become more specific and more numerous so the complexity of the task increases.

Table 5 also lists the average rate of substitutions as the annotation task's complexity increases, showing that the annotators were able to fairly easily distinguish between instances of the three subsystems of Appraisal (Attitude, Engagement and Graduation) as the substitution rate at level 1 is low (only 9%). As the number of possible classes increases annotators are more likely to confuse appraisal types, with disagreement occurring on approximately 44% of annotations at level 5. The second round of annotations resulted in slightly improved agreement at

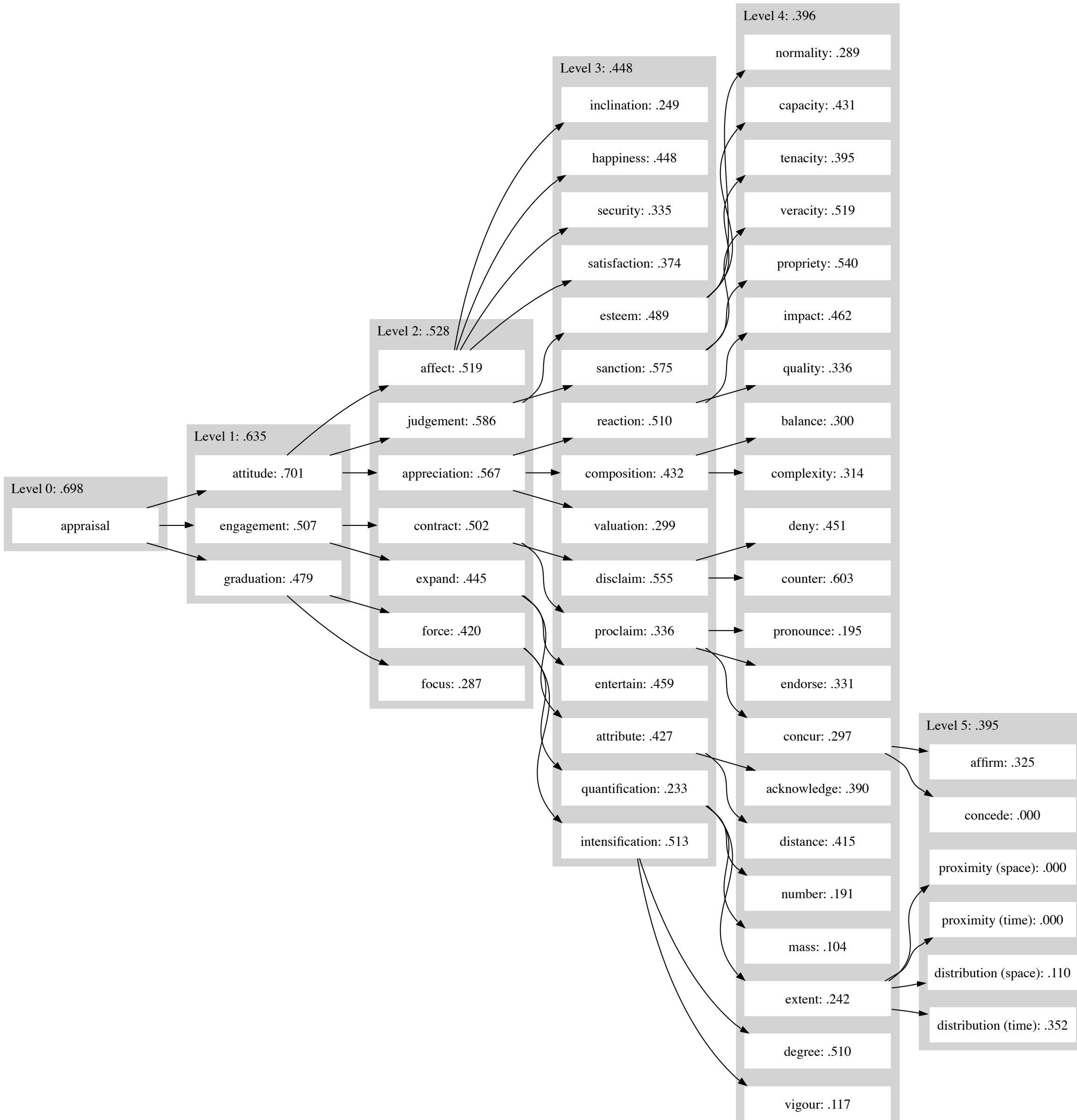

Figure 2: The Appraisal framework with hierarchical levels highlighted. Appraisal classes and levels are accompanied by the harmonic mean of the F-scores of the annotators for that class/level.

each level of abstraction (the mean F-score increased by 0.051 at the most abstract level).

Of course, some Appraisal classes are easier to identify than others. Figure 2 summarises the agreement for each node in the Appraisal hierarchy with the harmonic mean of the F-scores of the annotators for each class. Typically, the attitude annotations are easiest to identify, whereas the other subsystems of engagement and graduation tend to be more difficult.

The Proximity children of Extent exhibited no agreement whatsoever. This seems to have arisen from the differences in the judges' interpretations of proximity. In the case of Proximity (Space), for example, one judge annotated words that function to modify the spatial distance of other concepts (e.g. *near*), whereas the other selected words placing concepts at a specific location (e.g. *homegrown, local*). This confusion between modifying words and specific locations also accounts for the low agreement in the Distribution (Space) type.

The measures show that it is also difficult to achieve a consensus on what qualifies as engagements of the Pronounce type. Both annotators select expressions that assert the irrefutability of a proposition (e.g. *certainly* or *in fact* or *it has to be said*). Judge d, however, tends to perceive pronouncement as occurring wherever the author makes an assertion (e.g. *this is* or *there will be*). Judge j seems to require that the assertion carry a degree of emphasis to include a term in the Pronounce class.

The low agreement of the Mass graduations can also be explained in this way, as both d and j select strong expressions relating to size (e.g. *massive* or *scant*). Annotator j found additional but weaker terms like *largely* or *slightly*.

The Pronounce and Mass classes provide typical examples of the disagreement exhibited by the annotators. It is not that the judges have wildly different understandings of the system, but rather they disagree in the bounds of a class—one annotator may require a greater degree of strength of a term to warrant its inclusion in a class.

Contingency tables (not depicted due to space constraints) reveal some interesting tendencies for confusion between the two annotators. Approximately 33% of d's annotations of Proximity (Space) were ascribed as Capacity by j. The high percentage is due to the rarity of annotations of Proximity (Space), but the confusion comes from differing units of Appraisal, as shown in Example 6.

(6)
[d] But at key points in this story, one gets the feeling that the essential factors are operating *just outside*–PROXIMITY (SPACE) *James's field of vision*–CAPACITY.

[j] But at key points in this story, one gets the feeling that the essential factors are operating just *outside James's field of vision*–CAPACITY.

Another interesting case of frequent confusion is the pair of Satisfaction and Propriety. Though not closely related in the Attitude subsystem, j chooses Propriety for 21% of d's annotations of Satisfaction. The confusion is typified by Example 7, where it is apparent that there is disagreement in terms of *who* is being appraised.

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

[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.

Annotator d believes that the author is communicating the artist's dissatisfaction with the way he is treated by critics, whereas j believes that the critics are being reproached for their treatment of the artist. This highlights a problem with the coding scheme, which simplifies the task by assuming only one type of Appraisal is conveyed by each unit.

5 Related work

Taboada and Grieve (2004) initiated computational experimentation with the Appraisal framework, assigning adjectives into one of the three broad attitude classes. The authors apply SO-PMI-IR (Turney, 2002) to extract and determine the polarity of adjectives. They then use a variant of SO-PMI-IR to determine a 'potential' value for affect, judgement and appreciation, calculating the mutual information between the adjective and three pronoun-copular pairs: *I was* (affect); *he was* (judgement) and *it was* (appreciation). While the pairs seem compelling markers of the respective attitude types, they incorrectly assume that appraisals of affect are limited to the first person whilst judgements are made only of the third person. We can expect a high degree of overlap between the sets of documents retrieved by queries formed using these pairs (e.g. *I was a happy* $\langle X \rangle$; *he was a happy* $\langle X \rangle$; *It was a happy* $\langle X \rangle$).

Whitelaw et al. (2005) use the Appraisal framework to specify frames of sentiment. These "Appraisal Groups" are derived from aspects of Attitude and Graduation:

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

Their process begins with a semi-automatically constructed lexicon of these Appraisal groups, built using example terms from Martin and White (2005) as seeds into WordNet synsets. The frames supplement bag of words-based machine learning techniques for sentiment analysis and they achieve minor improvements over unigram features.

6 Summary

This paper has discussed the methodology of an exercise annotating book reviews according to the Appraisal framework, a functional linguistic theory of evaluation in English. The agreement exhibited by two human judges was measured by analogy with the evaluation employed for the MUC-7 shared tasks (Chinchor, 1998).

The agreement varied greatly depending on the level of abstraction in the Appraisal hierarchy (a mean F-score of 0.698 at the most abstract level through to 0.395 at the most concrete level). The agreement also depended on the type being annotated—there was more agreement evident for types of attitude compared to types of engagement or graduation.

The exercise is the first step in an ongoing study of approaches for the automatic analysis of expressions of Appraisal. The primary output of this work is a corpus of book reviews independently annotated with Appraisal types by two coders. Agreement was in general low, but if one assumes that the intersection of both sets of annotations contains reliable examples, this leaves 2,223 usable annotations.

Future work will employ these annotations to evaluate algorithms for the analysis of Appraisal, and investigate the usefulness of the Appraisal framework when in the computational analysis of document sentiment and subjectivity.

Acknowledgments

We would like to thank Bill Keller for advice when designing the annotation methodology. The work of the first author is supported by a UK EPSRC studentship.

References

M. M. Bakhtin. 1981. *The Dialogic Imagination*. University of Texas Press, Austin. Translated by C. Emerson & M. Holquist.

Rebecca Bruce and Janyce Wiebe. 1999. Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(1):1–16.

N. Chinchor. 1998. MUC-7 test scores introduction. In *Proceedings of the Seventh Message Understanding Conference*.

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

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

J. R. Martin and P. R. R. White. 2005. *Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

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

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 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.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. Longman.

M. Stubbs. 1996. Towards a modal grammar of English: a matter of prolonged fieldwork. In *Text and Corpus Analysis*. Blackwell, Oxford.

Maite Taboada and Jack Grieve. 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.

Peter D. Turney. 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.

P. R. R. White. 2002. Appraisal — the language of evaluation and stance. In Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen, editors, *Handbook of Pragmatics*, pages 1–27. John Benjamins, Amsterdam.

Casey Whitelaw, Navendu Garg, and Shlomo Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.