Linguistics and the Human Sciences

LHS (PRINT) ISSN 1742–2906
LHS (ONLINE) ISSN 1743–1662

Article

Contrastive analyses of evaluation in text: Key issues in the design of an annotation system for attitude applicable to consumer reviews in English and Spanish

Maite Taboada and Marta Carretero*

Abstract

This paper reports on part of the research on evaluative language currently carried out within the CONTRANOT project,[1] *which aims at the creation and validation of contrastive functional descriptions through corpus analysis and annotation in English and Spanish. More concretely, we will present the coding scheme designed for Attitude, a subcategory of Appraisal as studied within Systemic-Functional Linguistics (White, 2003; Martin and White, 2005). The criteria for selection and annotation of spans of Attitude in the coding scheme are specified and illustrated with examples from the Simon Fraser University Review Corpus (Taboada, 2008), a corpus of consumer-generated reviews on hotels, books and movies, and a small-scale English-Spanish contrastive analysis of these reviews has been carried out. The scheme is to be used for the future annotation of evaluation in an English-Spanish corpus, CONTRASTES (Lavid, 2008; Lavid* et al., *2007, 2010). Once annotated, the reviews will be part of this corpus.*

KEYWORDS: DISCOURSE ANALYSIS; EVALUATIVE MEANING; ANNOTATION SYSTEMS; CORPUS ANALYSIS

Affiliation

Simon Fraser University, Canada.
email: mtaboada@sfu.ca (corresponding author)

LHS VOL 6 2012 275–295 doi : 10.1558/lhs.v6i1-3.275

©2012, EQUINOX PUBLISHING

1. Introduction

Evaluative language, from the point of view of the Appraisal framework, refers to the linguistic expressions that indicate 'the subjective presence of writers/ speakers in texts as they adopt stances towards both the material they present and those with whom they communicate' (Martin and White, 2005: 1). The study of evaluative language has intrinsic interest, in that we all use language to evaluate, appraise and classify objects and people on an everyday basis. It has also received a surge of attention lately from more applied venues, in particular with regard to computational applications (see Pang and Lee, 2008, and references therein). The web is now teeming with opinions, which are of interest to marketers, policy makers and the public in general.

The field of sentiment extraction, or semantic orientation detection, is a growing area within computational linguistics. The approach typically taken consists of counting the number of positive and negative words in a text (usually adjectives), and averaging their values, determined by a pre-existing dictionary (e.g., Esuli and Sebastiani, 2006; Kennedy and Inkpen, 2006; Turney, 2002). Other approaches follow Machine Learning techniques, often involving little linguistic information (e.g., Pang *et al.*, 2002; Goldberg and Zhu, 2006). Some researchers have found the categories and classifications provided by the Appraisal framework of use, and are creating appraisal lexicons for this task (Taboada and Grieve, 2004; Whitelaw and Patrick, 2004; Whitelaw *et al.*, 2005; Bloom *et al.*, 2007). This proved efficiency of the Appraisal system for the creation of lexicons of evaluative expressions, together with the accessibility of its labels for use by non-linguists, are the reasons why we have adopted this theory for our analysis of evaluation. Among other approaches, we will quote Douglas Biber's approach, based on the quantitative analysis of clusters of grammatical and lexical features (Biber, 1988, 1995; Biber and Finegan, 1989a), and a few others which focus on the contribution of modality to express evaluation, stance or opinion in text (Stubbs, 1986; Biber and Finegan, 1989b) and the role of evidentiality in expressing subjectivity (Chafe and Nichols, 1986).

The Appraisal framework, which originated in Australia within the Systemic Functional School of Linguistics, was developed in response to the need to cope with (and, ideally, to be proficient in) the expression of interpersonal meaning. More concretely, Appraisal Theory was developed as part of the Disadvantaged Schools Program's *Write It Right* literacy project, which concerned writing in the workplace and secondary school (from 1990 to 1995 approximately). Its main proponents are Jim Martin and Peter White (see, for instance Martin, 2000; Martin and White, 2005, among other publications), but many other academics also participated (for example, Joan Rothery, Cate Poynton, Mary Macken-Horarik, Maree Stenglin, Rick Iedema and Susan Feez).

In Martin and White's words (2005: 9), Appraisal concerns 'how evaluation is established, amplified, targeted and sourced'.

The framework for the present study is a project which aims at developing contrastive corpus analyses, and then systems of annotations, for a number of linguistic categories, some of which (apart from evaluation, the focus of this paper) are coherence relations, tense, aspect and modality. The systems of annotation are to be designed for use by non-academic annotators, who will be provided with relatively simple sets of instructions. According to this aim, our approach to evaluation will have to be restrictive, giving preference to the individual evaluative charge of words and expressions against the overall evaluative effect of all the expressions in a given text. It could even be stated that in reality every word has its charge of evaluation: when we speak or write, even in those messages whose main role is to transmit information with a high degree of objectiveness, we design our utterances with the purpose of presenting a certain world view to the addressee.[2]

A related project focuses on detecting sentiment automatically, making use of linguistic information, and drawing on the insights provided by the Appraisal system. Preliminary work in the project has resulted in a collection of book, movie, and consumer product reviews (Taboada, 2008), a software program (Taboada *et al.*, 2008a), and a few publications and presentations (Taboada and Grieve, 2004; Taboada *et al.*, 2006a, b, 2008b, 2011; Voll and Taboada, 2007). This project is applied, seeking to develop an automatic system for the extraction of sentiment and evaluation in text. The work described in this paper is more theoretical, in the sense that the goal is to understand evaluation in text, and to provide an annotated corpus. This work will also draw on previous work on modality and its role in the expression of the speaker/writer's attitude (Carretero, 2002, 2004, 2007).

The goal, then, was to create a system for annotating evaluation in text. The first step involved a detailed analysis, based on a corpus of authentic texts. In this way, a wide variety of patterns and relations that convey evaluation can be found, and are labelled for analysis. For its purposes, the analysis has to meet two characteristics: (1) rigour and clarity, so as to ensure inter-rater reliability; (2) simplicity, with the view that this analysis will be the basis of an annotation system designed for use by non-specialists.

2. The corpus and its annotation

The corpus selected for analysis is part of the larger Simon Fraser University Review Corpus, which consists of 1,600 movie, book, and consumer product reviews, 800 in English and 800 in Spanish. The English reviews were extracted from the web page Epinions.com. A first data collection (400 texts) took place in 2004, and a second round of the same number was collected in 2008. For Spanish, we used two web sites: Ciao.es and Dooyoo.es, all of it collected in 2008. The reason to have two rounds for each language was that we used one set to develop dictionaries and other resources, and thus we needed a set for independent evaluation purposes. The reviews are divided into eight categories: books, cars, computers, cookware/appliances, hotels, movies, music and phones.

The reviews are all written by non-experts. The contributors are mostly avid web users who enjoy sharing their experiences with others. In addition, Ciao promises a small monetary compensation if a sufficiently large number of reviews are posted and they are considered useful by readers of the site. The language is typically informal, with many colloquial expressions, typos and abbreviations. In the examples provided in the paper, we have left punctuation, spelling and grammar as they originally appeared.

3. Categories within the Appraisal system

Within the Appraisal system, the types of evaluation are divided into three broad categories: Attitude, Graduation and Engagement. The main proponents of the Appraisal System, James R. Martin and Peter R. R. White, have acknowledged that these labels are not unarguable; for example, White (2002: 7), writing about Attitude and its subtypes, states that they have been designed 'as a resource for those who need something to manage the analysis of evaluation in discourse, and as a challenge to those concerned with developing appropriate reasoning'. Appraisal in fact has been criticized for the arbitrariness of the labels and the difficulties that it poses for inter-rater reliability. However, we believe that it is still possible to design a system that guarantees this kind of reliability, even if some of the decisions made will unavoidably have some degree of arbitrariness.

3.1. Attitude

Attitude concerns the expression of feeling, and is subdivided into three types:

- **Affect**, which covers the explicit expression of positive or negative feelings by the speaker/writer or someone else, as in *I am* ***happy****, She* ***likes*** *him, He left the office* ***sadly****.*
- **Judgement**, which 'deals with attitudes towards behaviour, which we admire or criticise, praise or condemn' (Martin and White, 2005: 42). Judgement concerns social esteem and ethical evaluations, and applies mostly to persons or institutions. Examples of this category are *They acted* ***honestly*** or *She is an* ***efficient*** *worker.*

- **Appreciation**, which ‘involves evaluations of semiotic and natural phenomena, according to the ways in which they are valued or not in a given field’ (Martin and White, 2005: 43). Examples of Appreciation are *This book is **fascinating*** or *The plot is **conventional***. The evaluations are aesthetic or functional, and they refer mostly to works of art or literature and to non-human physical objects, rather than to humans.

From the description above it may be inferred that Affect evaluates the entity through the expression of feeling (the speaker/writer’s, or that of someone else), whereas Judgement and Appreciation evaluate the entity by attributing a quality to it.

Among all three categories, Attitude will be the main concern of this paper, and at the first level of analysis. Each of the Attitude categories, Affect, Judgement and Appreciation, is divided into subcategories, but these will not be taken into account for the sake of simplicity.

3.2. Graduation

Graduation consists of the use of linguistic expressions for emphasizing or downtoning other expressions. Expressions of Graduation differ from those of Attitude in that they do not have intrinsic positive or negative values by themselves, but acquire them in context. Some expressions of Graduation are intensifiers applied to nouns (*real, true, genuine*) or to adjectives (*very*, *really*), and softeners (*kind of, sort of, or something*). Graduation is divided into two broad subtypes: Focus and Force.

- **Focus** involves Graduation according to prototypicality, that is, ‘by reference to the degree to which they match some supposed core or exemplary instance of a semantic category’ (Martin and White, 2005: 137). Focus is divided into the subtypes **Sharpen** and **Soften**, which indicate proximity and distance, respectively, to a core or exemplary member of the category. Some examples of expressions realizing Sharpen are *real, true, genuine(ly), effective(ly)* … and instances of Soften are *kind of, sort of, of sorts, -ish (fourish), bordering on* …
- **Force** serves speakers or writers to modulate the impact of what they say. Force is divided into the following categories:
    - **Intensification**, which can apply to a quality (*slightly sad*) or to a process (*greatly disturbed me*), but no difference will be made in this respect concerning annotation. Some realizations of intensification are *a bit, somewhat, relatively, fairly, rather, very, extremely, utterly;* self-pronouns when their use is optional (1); the comparative and superlative forms or constructions with adjectives. Intensification may also apply to an entity, as in (2):

(1) He did it *himself.*

(2) This is the *very* book I was reading the other day.

equinoxonline

- **Quantification**, which is divided into the following categories:
    - Number: *a few, lots of, many, streams of*. Exact numbers are not considered to be Appraisal devices, since they refer to objective quantities.
    - Mass/presence: *tiny, small, large, huge, gigantic*
    - Extent:
        - Proximity:
            - Time: *recent, ancient*
            - Space: *nearby, distant*
        - Distribution:
            - Time: *long-lasting, fast*
            - Space: *narrow, broad*

In this paper, Graduation will be considered only when it is within the scope of an evaluative span of Attitude. For example, in (3), the evaluative span is *a little excited*: the key expression is *excited* (Affect), which is modified by the expression of Graduation *a little*:

(3) I got **a little excited** that things were looking up only to find out that it really was nothing. (Books no, 3)

3.3. Engagement

Engagement 'deals with sourcing attitudes and the play of voices around opinions in discourse' (Martin and White, 2005: 35). Examples of Engagement are epistemic modal expressions (*He* ***might*** *have finished his studies by now*), evidential expressions (***Apparently****, he has recovered from his illness*) or denials (*This hotel* ***is not*** *near the sea as you said*). Engagement will only be marginally discussed in borderline cases with Attitude (see 5.6).

4. The experiments

For a qualitative analysis, we did a first experiment with 12 reviews, distributed evenly according to these features: language (six English, six Spanish); kind of evaluation (six positive, six negative); product evaluated (four books, four movies, four hotels). After the reviews had been analysed in terms of problematic issues and the results discussed, we designed a second experiment. This time we restricted the experiment to books and movies which, in contrast to hotels, are intellectual products that have authors, plot and characters. The analysis took place in two steps: the first was the selection of the markables; once agreement was reached, we undertook the labelling of these markables. For this second experiment, we selected eight reviews, four for each language, equally divided between reviews of books and movies, and positive and negative reviews. We found that selecting the markables was the most difficult

equinoxonline

task, whereas labelling them was easier, and led to higher agreement. Table 1 shows the results for the selection of markables for the eight reviews. We see that the initial total agreement is quite high. The precision (calculated as the number of units for the annotator with the lowest number of units, Annotator 2, over the number of units for Annotator 1) is quite high. More importantly, both total and partial recall (number of agreements over the number of units for Annotator 1) is very high.

Table 1: Agreement for annotation experiment

Units (Annotator 1)	348
Units (Annotator 2)	315
Total agreement	281
Partial agreement	31
No agreement	47
Recall	90.52%
Precision (total)	80.75%
Precision (partial)	89.66%

Once we were confident that our annotations were reliable and showed high degree of agreement, we proceeded to annotate further texts. The entire corpus discussed in this article (see Section 2) consists of 32 reviews, 16 in each language. Throughout the paper, we will refer to examples from the corpus; occasionally, examples from other reviews will also be used. When examples from this corpus are used, the product reviewed and the number of the review will be specified. Non-labelled examples are constructed by the authors.

5. Selection of markables of Attitude

5.1. Application of the evaluation to the products evaluated

Evaluation was restricted to the cases in which it refers to the products evaluated and related entities that reflect the quality of the product, such as the author, the plot and quality features of the character of books (*believable, deep …*), or performances of actors and actresses in movies. We have excluded evaluative spans included in the descriptions of the plot or characters in books and movies: a director may well have chosen an ugly suburb of a city as scenario or a stupid person as the protagonist so as to suit best his/her purposes. Examples of non-included evaluative spans are (4) and (5). However, trailers and covers of books were considered as part of the movie or book; consequently, the evaluative expressions referring to them were included (6).

equinoxonline

(4) The Spruills have a son who is a rather large bully (Books no, 3)

(5) This book was about a lawyer who worked in a firm as a litigator and was around the most defiant and high up men there were in the city. (Books no, 1)

(6) The trailer did manage to make the film look fun (Movies yes, 1)

5.2. Evaluative and non-evaluative occurrences of the same word or expression

Some words or expressions have an inherent evaluative meaning, so that they are always considered as cases of Appraisal. These include many adjectives (*disastrous, excellent, fair, great …*) and their derived words. Among other kinds of evaluative words other than adjectives and their derived words, there are quantifiers, such as *too (much)*, which is negative in that indicates excess; nouns (*joy, sorrow*) or verbs (*excel, improve, disappoint*). This is also the case of some grammatical constructions, such as *all that* and its Spanish correlate *todo ello*, literally 'it all' or *no hace más que* 'does nothing but', which have an evaluative meaning (negative in these cases).

Concerning expressions that indicate manner, Martin and White (2005: 146) follow Stillar's (1998) argument that circumstances of manner (7) always implicate the speaker/writer's subjectivity, since they do not lend themselves to objective accounts in the same way as time, place or cause do. Accordingly, we have always included these expressions as evaluative spans.

(7) she comes out *very shrilly*. (Movies no, 1)

Other items, however, may be considered evaluative or not, depending on the context. Some examples from the reviews in which the expressions can be considered as evaluative due to the context are (8) and (9). In (8), *típica* 'typical' has negative connotations, but this adjective can be easily imagined in a neutral context ('the typical costume of the village'). In (9), *generic* has the sense of 'clichéd, stereotyped' and has therefore been considered as the head of an Appraisal span; however, it has non-evaluative uses, as in *generic software*, which means software for a wide range of computers.

(8) A la protagonista nos la presentan como a la *típica* mujer que sabe que consigue más luciendo carne, que utilizando el cerebro, en fín en una palabra decepcionante. (Libros no, 1.14).

'The protagonist is introduced as the *typical* woman who knows that she achieves more showing her flesh than using her brain, to sum up, in a word disappointing.'

(9) her character was *just totally generic* (Movies no, 1)

5.3. Invoked evaluations

In contrast to cases where the evaluation is due to the lexical meaning of a word or expression (which is called Inscribed evaluation), in other cases there are facts that imply positive or negative evaluation. In those cases, the evaluation is Invoked. In our annotation system, spans that could possibly be considered as invoked evaluations have been excluded in many cases, for the sake of simplicity. For example, (10) is part of the argument that the treatment of the role of women in the 1950s is inaccurate, and deviates from what was really the case. The reviewer suggests that this lack of authenticity may be due to the audience to which the film was addressed. Similarly, (11) refers to a fact (*hacer taquilla*, which literally means 'make box office') which may well be considered as morally questionable. As we stated above, these cases will not be included in our evaluation analysis.

(10) I know women were the core audience of this film, particularly young college women. (Movies no, 1)

(11) Steven está claro a lo que ha ido que es a hacer taquilla (Películas no, 1.11)

'It's clear that what Steven wanted was big box-office numbers.'

However, we have included invoked evaluations in a number of cases, in which the linguistic clues facilitate the consideration as such:

1. **Complex clauses**, one of which suggests (not) to read or view the product evaluated and the other provides argumentation for this suggestion (*if*- conditionals or similar constructions). In these cases the evaluation is not communicated by lexical meaning, but by implicature. Both (12) and (13) are spans with an implicated negative evaluation, the book reviewed being *The Da Vinci Code*. The evaluation is even more indirect in (13), since it mentions other novels, hinting that the reviewed book should not be read.

(12) For an example of how marketing hype can overcome critical judgment and influence popular taste, read 'The Da Vinci Code'. (Books no, 1)

(13) If you're looking for an intellectually challenging mystery story, read or reread Eco's 'The Name of the Rose', or 'Foucault's Pendulum'. (Books no, 1)

2. When **comparison** is used for evaluative purposes. An example is (14): Although the phrase 'like Sherlock Holmes on speed' is not negative *per se*, it is here used as criticism for the hectic pace of the book. In certain cases, the comparison is not with entities, but with situations (15):

(14) Well actually, you have until Saturday Night so there's time to run around *like Sherlock Holmes on speed* and solve the mystery just minutes before the news media puts your company out of business. (Books no, 19)

(15) Descriptions of places – Louvre, Westminster Abbey – are lifeless, and read as if plagiarized from a do-it-yourself walking tour guide by one of the less gifted of the author's former prep school students. (Books no, 1)

3. When **metaphors** are used for evaluative reasons. We have to specify that our approach to metaphor is more restricted than that commonly used in cognitive linguistics (e.g., Lakoff and Johnson, 1980; Coulson, 2001; Fauconnier and Turner, 2008), according to which many cases of transfer of domain are considered as metaphors. For example, *go into* in (16) would be considered as metaphorical, since it does not express physical movement. We will consider as metaphors only those cases in which the writer clearly has consciousness of this transfer of domain. One such example is (17), in which *Madame Bovary* is an expression of evaluation (and the ensuing paraphrase gives a clue of the sense of this evaluation).

(16) He went into the matter carefully. (Cf. 'He went into the labyrinth carefully')

(17) No solemos dudar del amor que nuestra madre nos profesa ni del que tenemos a nuestros hijos o amigos, y sin embargo, no necesitamos reafirmarlo con expresiones verbales. […] Sin embargo ¿por qué sí lo esperamos de nuestras parejas? Yo intuyo que es por una necesidad creada por la literatura (en su conjunto) de ser *Madame Bovary*, es decir, la protagonista de nuestra propia novelita rosa. (Libros yes – 4.11.)

'We don't usually doubt the love that our mother has for us nor that we have for our children or friends, and nevertheless, we do not need to reaffirm it with verbal expression. […] However, why do we expect that from our lovers? I believe that it is because of a need created by literature (as a whole) to be *Madame Bovary*, that is, the protagonist of our own little romance novel.'

5.4. Emotional outbursts and vocatives

Martin and White (2005: 68) classify swearing as Involvement, a meaning which, like Appraisal, is included within Tenor, and concerns the distinction between proximal and distal stance towards the text and the addressee. We believe that in our corpus these expressions (*my God, hell,* etc.) as well as vocatives (*honey, my dear, idiot, son of a bitch …*) have, above all, an evaluative role of Affect, since they express strong positive or negative feelings. The same may be stated about emotional outbursts different from swearing: In (18) *ooooh* intensifies the absurdity of the conspiracy, and therefore could be considered as a span of negative Appreciation, and in (19) the initial outburst lays emphasis on the bad quality of the film:

(18) *Ooooh* ... big conspiracy ... it would be nice if James Patterson explain why. (Books no, 24)

(19) *Bufffffffffff*, por donde empiezo? (Películas no, 2-10)
'Bufffffffffff, where do I start?'

5.5. Inclusion of markables in elliptical expressions or expressions replaced by a pronoun

We have included the spans in which an evaluative word or expression is inferable from the linguistic context and omitted by ellipsis (20) or substitution (21).

(20) I guess I just thought that this movie would be as good as the Grinch, but unfortunately, it *wasn't*. (Movies no, 13)

(21) tenía cierto miedo a que mis queridos Simpson perdieran en su aparición en la pantalla grande. Pero *no ha sido así*. (Películas yes, 4-2)

'I was quite afraid that my dear Simpsons would lose in their appearance on the big screen. But it hasn't been so.'

5.6. Overlaps between Attitude and Engagement

There are some expressions that overlap between Attitude and Engagement. Negative or non-assertive linguistic devices are a subtype of Engagement, in the sense that the contents communicated (have the potential to) clash with previous expectations, and will be considered as such in our annotation scheme, but at the same time they point to evaluations (22–23):

(22) she *doesn't* really bring *anything* that we *haven't* seen before (Movies no, 1)

(23) the reader *still* has gotten *few* clues about matters (Books no, 1)

Another case in point are epistemic and deontic modal expressions. Epistemic modality, which may be defined as the estimation of the chances that a state of affairs has of being or becoming true, has a high degree of overlap with Attitude (Carretero and Taboada, to appear). We have classified these cases within Attitude or within Engagement, depending on the relative importance of the emotional or the epistemic meaning. Expressions of credibility (24), sincerity (25), (ab)normality (26), and (dis)agreement with expectations (27) are classified under Attitude, while those of probability due to quality (28) are classified under Engagement.

(24) One of the few positives of the film is the cinematography by Anastas N. Michos, that has lovely moments of colorful images and that *authentic* 1950s look. (Movies no, 1; Appreciation)

(25) *Sinceramente* me esperaba que fuese peor. (Películas yes, 5–11)

'*Frankly*, I expected it to be worse.'

(26) Dos policías. – el más joven intenta imponer su ley ante la vecindad desesperada. *Lógicamente*, NO lo conseguirá. (Películas no, 1–5)

'Two policemen.- the youngest one tries to impose his law on the despairing neighbourhood. *Obviously***,** he WON'T manage it.'

(27) el filme *no me ha decepcionado en absoluto* (Películas yes, 4–2)

'The movie did not disappoint me at all'

(28) it's *likely* not to get any Oscar nominations for anyone involved except the costume and production designs. (Movies no, 1)

Deontic modality, that is, obligation, recommendation and permission, is characterized in Martin and White (2005: 111) in terms of Engagement, with the argument that this modality 'explicitly grounds the demand in the subjectivity of the speaker – as an assessment by the speaker of obligation [or of permission] rather than as a command'. However, we believe that these expressions have an important semantic feature of Judgement: Obligations and recommendations, as in (29), are morally desirable events, and permissions are morally acceptable events. Therefore, we will classify these expressions under Judgement.

(29) Newell still *should be given* credit for trying to make things interesting since the pacing of the film is attentive but it's given a weak script with no sense of a singular direction to begin with. (Movies no, 1)

Rhetorical questions are also devices that express Attitude, but we believe that their main meaning belongs to Engagement, in that they have a strong implicature of positive or negative polarity. For example, the evaluative span in (30) has been classified under Engagement, but could also be considered as negative Appreciation, since the reviewer criticizes an inconsistency of the book.

(30) …While talking about conspiracies involving the Whitehouse, James Patterson mentions the Whitewater Scandal. *How can that be if Clinton was never the President in Patterson's story?* (Books no, 24)

5.7. Inclusion of expressions of Graduation within spans of Attitude

In some cases, the evaluative expression of Attitude is intensified or downtoned by a word or expression of Graduation that syntactically modifies it. In these cases, the modifier is included in the evaluative span (31–32). We will only annotate the realizations of Graduation by independent words: our approach will not consider scalar terms, as in the series of terms *contented/happy/joyous* (Affect), *competently/skilfully/brilliantly* (Judgement) or *warm/hot/scalding* (Appreciation). On the other hand, we will adopt Martin and White's (2005: 143) inclusion under Graduation of the expressions with a lexical meaning of Attitude, such as *reasonably happy* or *dreadfully cold*, on the grounds that their effect in these contexts is to intensify the meaning of the evaluative expression that they modify. In other words, they undergo delexicalization, even though it might be argued that their meaning of Attitude is not entirely lost.

(31) There is just something about the way he says his lines that makes them *so funny*. (Movies no, 13)

(32) That's probably *the biggest detriment* to the book (Books no, 17)

5.8. Length of the text spans

Due to the overall aim of the analysis (to annotate a large quantity of text spans), we restrict the spans to the evaluative lexical item, leaving aside the constituents of the syntactic unit to which it belongs. For example, in (33) the span is restricted to *free* instead of the whole constituent (*free parking for hotel guests*), and in (34), the span includes only *decrepitud* 'decrepitude', rather than *la decrepitud de Ender*.

(33) I would definitely recommend the Golden Nugget (oh, and did I mention *free* parking for hotel guests?) (Hotels yes, 22)

(34) la *decrepitud* de Ender (Libros no, 1.1)
'Ender's decrepitude'

However, some spans consist of more than one word, since the kind of evaluation that they express depends on the expression as a whole. In (35) the span is *bajando la calidad*, since neither *bajando* nor *la calidad* convey the negative evaluation expressed by the whole:

(35) y después fue *bajando la calidad* hasta llegar a su final (Libros no, 1.1)
'and afterwards *quality kept going down* until it reached the end'

In certain cases, especially when the expressions of Attitude are modified by expressions of Graduation, the evaluative spans are discontinuous, with non-evaluative items in the middle. However, the non-evaluative words have been included within the spans for reasons of easiness of quantification:

(36) *lo peor* [*que he leído*] *en mucho tiempo* (Libros no, 1. 11)
'the worst that I have read in a long time'

5.9. Coordinated and juxtaposed evaluative expressions

When evaluative expressions are joined by a coordinating conjunction, they are considered as a single span, since they can be the scope of a single Graduation

equinoxonline

expression (37). This is not the case of juxtaposed spans, which are consequently considered as separate spans (38). However, when two coordinated spans are modified by different expressions of Graduation, they are considered as separate spans (39).

(37) Julia Navarro ha conseguido que la historia sea *interesante y apasionante desde el principio hasta el final*. (Libros yes, 5–10)

'Julia Navarro has managed to make the story interesting and exciting from beginning to end'

(38) It's a *brazen/daring/no-holds-barred comic* assault on many of the values that we hold most dear (Movies yes, 23)

(39) I also found Block's transformation from a money-hungry associate striving to make partner in a large firm to an idealistic lawyer hoping to change the world *a bit forced* and *somewhat unconvincing*. (Books no, 11)

6. Labelling of the markables

6.1. Criteria for signalling subcategories within Attitude

Ethics and aesthetics

In order to annotate the spans as instances of Judgement or Appreciation, the first distinction to consider is that between **ethics** and **aesthetics**. Evaluations about ethics are under Judgement, and evaluations about aesthetics are under Appreciation, independently of whether the target is human or non-human: in this way, *a fair referee* and *a fair decision* are both classified under Judgement, while *an ugly dress* and *an ugly person* are both cases of Appreciation.

Human and non-human targets

When the evaluation cannot be easily categorized into ethics or aesthetics, it is classified under Judgement if the target is human (*an efficient actor*) and otherwise as Appreciation (*an efficient computer*). However, entities named by abstract nouns are classified as Judgement, since they are nominalizations of the actions of persons or institutions. For example, (40) could equally be expressed by 'the publisher worked unusually hard on the marketing':

(40) The success of the book must be attributed to the publisher's (Doubleday) *unprecedented marketing effort*. (Books no, 1)

The influence of the context

Some lexical items are associated with Judgement, and others with Appreciation. However, lexical items normally associated with one of these categories can occasionally realize the other. For example, the adjective *stupid* is associated with Judgement, but realizes Appreciation in *a stupid novel.* According to Bednarek (2009: 182), these instances provoke the effect of 'a collocational clash and a particular flavour of appraisal meaning'. For example, the periphrastic construction with *poder* 'can, be able' with a human subject is normally associated with capacity (Judgement), but in (41) the determining factor for this capacity does not lie in the value of children, but in the value of the movie, and hence the span has been classified as Appreciation.

(41) Los peques *podrán aprender*, gracias a Wall-e, el valor de cuidar el planeta para que nos dure un poquito más. (Movies yes, 5–10)

'Kids will be *able to learn*, thanks to Wall-e, the value of taking care of the planet so that it will last a little longer.'

Another point is the distinction between Affect and Judgement in terms such as *guilty, embarrassed, proud, jealous, envious, ashamed …* (Martin and White, 2005: 60). We classify these expressions under Affect when they express feelings, i.e., *John is jealous of some of his wife's male colleagues*, and under Judgement when they express character traits of individuals, as in *John is a jealous person*.

Adjectives of reaction

Some adjectives indicate the emotions that the entity in question provokes in the reader. However, the adjective conceptualizes this emotion-triggering as a quality, so that they are classified under Judgement or Appreciation, not as Affect:

(42) Even in the tense, dramatic moments, it doesn't feel *suspenseful* nor in the lighter moments, it comes out as fluff. (Movies no, 1)

(43) Unfortunately, isolated examples can't create the *mind-numbing* effect of page after page of this tedious bloviating. (Books no, 1)

6.2. Polarity

Within the Appraisal framework, positive and negative polarity are associated with favourable and unfavourable evaluations, as in Examples (44) and (45), respectively.

(44) Janet Evanovich's series of books starring Stephanie Plum, an inept bounty hunter, was one of the *most enjoyable* books I've read in a while. (Books yes, 3)

(45) What *really put me off* was that it was *not clean* (hairs in the tub, dust in the mini-bar, etc.) (Hotels no, 1–11)

In the examples listed above, polarity is determined by lexical meaning. In other cases, it depends heavily on context. In (46) the negative polarity is due to the entity evaluated: if it had been a medicine or curative plant instead of a book, irony would be out of place and the polarity would be positive. In (47), the counterfactual conditional reverses the polarity. In (48), the positive lexical item occurs in a comparison with other related entities, so that the polarity of the span is negative.

(46) contra el insomnio es *infalible* (Libros no, 1.14).
'against insomnia it is *infallible*'

(47) si tuviera un *buen* sumario (Libros no, 1.12)
'if it had a *good* index'

(48) I just felt her other novels were *much more exciting and interesting*. (Books no, 17)

7. Comparison of Appraisal realization across languages

The annotations were carried out on 32 texts, 16 per language, for a total number of words of 11,990 in English and 19,507 in Spanish. This being such a small corpus, we cannot make any broad generalizations about the type of phenomena found across the two languages. Here, we will merely show some differences that seem to be developing as trends in the corpus.

The two corpora contained different numbers of tokens of Attitude: 237 in English (one per 50.59 words) and 687 in Spanish (one per 28.39 words). In spite of this quantitative difference, when it came to distribution by Attitude type, the percentages were quite similar, with Appreciation having the overwhelming majority of the tokens, and Affect and Judgement more or less sharing the balance. In terms of polarity, the two languages also show similar trends, with a majority of positive tokens, albeit Spanish has a wider gap between positive and negative tokens. Table 2 summarizes these statistics.

Table 2: Feature statistics for the corpus

	English		Spanish	
	Tokens	Percentage	Tokens	Percentage
Attitude (tokens)	237		687	
Affect	45	18.99%	136	19.80%
Judgement	52	21.94%	168	24.45%
Appreciation	140	59.07%	383	55.75%
Positive polarity	121	51.05%	425	61.86%
Negative polarity	116	48.95%	262	38.14%

equinoxonline

8. Conclusions and suggestions for future research

In this paper we have described the scheme for Attitude within the CONTRANOT project, based on two experiments, the first of which included annotation of consumer reviews of books, movies and hotels and the second was restricted to books and movies. Concerning the selection of the evaluative spans, the signalling is restricted to the cases in which it refers to the products evaluated and related entities, trailers of movies and back covers of books. The words and expressions that have an inherent evaluative meaning are therefore systematically included in spans, while others are only evaluative in certain contexts. Invoked evaluations (that is, evaluations by implicature), are mainly comparisons and metaphors, as well as complex clauses containing a suggestion about the product reviewed as well as argumentation for this suggestion. Other kinds of evaluative spans are emotional outbursts and vocatives, spans containing omitted evaluative lexical items by ellipsis or substitution and expressions of deontic modality. Some areas of overlap between Attitude and Engagement, such as epistemic modality, rhetorical questions and some instances of negation and non-assertion, have been discussed with regard to the selection of the Attitude spans. As for the length of the text spans, the tendency is to restrict it as far as possible. An exception to this tendency is the inclusion of expressions of Graduation within the scope of those of Attitude. Evaluative lexical items joined by coordinating conjunctions are considered as a single span, but not if they are juxtaposed.

With regard to the annotation of markables, the key criteria for signalling subcategories are ethics and aesthetics, as well as human and non-human targets. Some lexical items are mainly associated with Judgement and others with Appreciation, but their value may vary depending on context. Concerning polarity, the main perspective adopted is that of the entity reviewed, so that in certain cases the positive or negative value of the span is the opposite of its lexical meaning.

Due to the complexities involved in the design of a coding scheme for Attitude, the two experiments were necessary so as to reach a satisfactory degree of agreement between the annotators. Even though we cannot make broad generalizations due to the size of the corpus, we should state that the percentages of the three subtypes of Attitude (Affect, Judgement and Appreciation) were similar in the English and in Spanish reviews; differences were found in the number of spans, which was higher in Spanish both in absolute terms and in frequency per number of words, and also in polarity, in that positive polarity displayed a higher percentage in the Spanish reviews while the opposite occurred with negative polarity.

This scheme for Attitude could be further refined by analysing a higher number of consumer reviews and by diversifying the kinds of products, in order to arrive at an easily reproducible and transparent standard of annotation.

Notes

[1] The CONTRANOT project is financed by the Spanish Ministry of Science and Innovation under the I+D Research Projects Programme (reference number FFI2008-03384). As members of the team, we gratefully acknowledge the support provided by Spanish Ministry and also the BSCH-UCM grant awarded to our research group.

[2] The system has been designed considering a discussion on the Appraisal Analysis e-mail list initiated by Marta Carretero, which took place in January 2010. We thank Monika Bednarek, Geoff Thompson, Alexanne Don and Donna Miller for their contributions. The shortcomings and inconsistencies of the resulting system are our responsibility.

About the authors

Maite Taboada is Associate Professor of Linguistics at Simon Fraser University (Canada). She holds MA and PhD degrees from the Universidad Complutense de Madrid (Spain), and an MSc in Computational Linguistics from Carnegie Mellon University. Maite works in the areas of discourse analysis, systemic functional linguistics and computational linguistics.

Marta Carretero is Associate Professor in English Language and Linguistics at the Universidad Complutense, Madrid, where she currently lectures on the areas of pragmatics and functional linguistics. She has done extensive research on modality, including English-Spanish contrastive analysis, the influence of genre on modality and the relationship between modality and evidentiality, among other issues. Within this area she authors articles and reviews in international and Spanish periodical journals. She also has contributions in collective works, such as English modality in perspective: Genre analysis and contrastive studies (2004), *Perspectives on Evidentiality and Modality* (2004) and *Studies on English Modality: In Honour of Frank Palmer* (2009). She has directed two research projects on modality in English and Spanish. She is co-editor of the collective work *A Pleasure of Life in Words: A Festschrift for Angela Downing* (2006), and was managing editor of the academic journal *Atlantis* (2006–2008).

References

Bednarek, Monika (2009) Language patterns and ATTITUDE. *Functions of Language* 16 (2): 165–192. http://dx.doi.org/10.1075/fol.16.2.01bed

Biber, Douglas (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511621024

equinoxonline

Biber, Douglas (1995) *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511519871

Biber, Douglas and Finegan, Edward (1989a) Drift and the evolution of English style: A history of three genres. *Language* 65 (3): 487–517. http://dx.doi.org/10.2307/415220

Biber, Douglas and Finegan, Edward (1989b) Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9.1: 93–124. http://dx.doi.org/10.1515/text.1.1989.9.1.93

Bloom, K., Garg, N. and Argamon, S. (2007) Extracting appraisal expressions, *Proceedings of HLT/NAACL* 308–315. Rochester, NY.

Carretero, Marta (2002) The influence of genre and register on epistemic modality in spoken English: A preliminary study. *Estudios Ingleses de la Universidad Complutense* 10: 11–41.

Carretero, Marta (2004) The role of evidentiality and epistemic modality in three English spoken texts from legal proceedings. In J. Marín-Arrese (ed.) *Perspectives on Evidentiality and Modality in English and Spanish,* 25–62. Madrid: Editorial Complutense.

Carretero, Marta (2007) Subjectivity in English epistemic modality: A two-resource based approach. *BELL New Series* 5 (5): 97–111.

Carretero, Marta and Taboada, Maite (to appear) The annotation of Appraisal: How attitude and epistemic modality overlap in English and Spanish consumer reviews. In J. R. Zamorano (ed.) *Thinking Modally: English and Contrastive Studies on Modality*. Berne: Peter Lang.

Chafe, Wallace and Nichols, Johanna (1986) *Evidentiality: The Linguistic Coding of Epistemology*. Norwood, NJ: Ablex.

Coulson, Seana (2001) *Semantic Leaps: Frame-shifting and Conceptual Blending in Meaning Construction.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511551352

Esuli, Andrea and Sebastiani, Fabrizio (2006) Determining term subjectivity and term orientation for opinion mining, *Proceedings of EACL-06, the 11th Conference of the European Chapter of the Association for Computational Linguistics*. Trento, Italy.

Fauconnier, Gilles and Turner, Mark (2008) Rethinking metaphor. In R. Gibbs (ed.), *Cambridge Handbook of Metaphor and Thought,* 53–66. Cambridge: Cambridge University Press.

Goldberg, Andrew B. and Zhu, Xiaojin (2006) Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization, *Proceedings of HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing,* 45–52. New York.

Kennedy, Alistair and Inkpen, Diana (2006) Sentiment classification of movie and product reviews using contextual valence shifters. *Computational Intelligence* 22 (2): 110–125. http://dx.doi.org/10.1111/j.1467-8640.2006.00277.x

Lakoff, George and Johnson, Mark (1980) *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Lavid, Julia (2008) CONTRASTES: An online English-Spanish textual database for contrastive and translation learning.

Lavid, J., Arús, J. and Zamorano, J. R. (2007) *Working with a Bilingual English-Spanish Database using SFL*. Paper presented at the 34th International Systemic-Functional Congress, Odense, Denmark.

Lavid, J., Arús, J. and Zamorano, J. R. (2010) *Towards an Annotated English-Spanish Corpus with SFL Textual Features*. Paper presented at the 37th International Systemic-Functional Congress, Vancouver, Canada.

Martin, James R. (2000) Beyond exchange: Appraisal systems in English. In S. Hunston and G. Thompson (eds) *Evaluation in Text: Authorial Distance and the Construction of Discourse*, 142–175. Oxford: Oxford University Press.

Martin, James R. and White, Peter R. R. (2005) *The Language of Evaluation*. New York: Palgrave.

Pang, Bo and Lee, Lillian (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2 (1–2): 1–135. http://dx.doi.org/10.1561/1500000011

Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up? Sentiment classification using Machine Learning techniques, *Proceedings of Conference on Empirical Methods in NLP*, 79–86.

Stillar, Glenn (1998) *Analyzing Everyday Texts: Discourse, Rhetoric and Social Perspectives*. London: Sage.

Stubbs, Michael (1986) 'A matter of prolonged field work': Notes towards a modal grammar of English. *Applied Linguistics*, 7 (1): 1–25. http://dx.doi.org/10.1093/applin/7.1.1

Taboada, Maite (2008) SFU Review Corpus [Corpus]. Vancouver: Simon Fraser University, http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html.

Taboada, Maite, Anthony, C., Brooke, J., Grieve, J. and Voll, K. (2008a) *SO-CAL: Semantic Orientation CALculator*. Vancouver: Simon Fraser University.

Taboada, Maite, Anthony, C. and Voll, K. (2006a) Creating semantic orientation dictionaries, *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)* 427–432. Genoa, Italy.

Taboada, M. Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011) Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37 (2): 267–307. http://dx.doi.org/10.1162/COLI_a_00049

Taboada, M., Gillies, M. A., and McFetridge, Paul (2006b) Sentiment classification techniques for tracking literary reputation, *Proceedings of LREC Workshop: Towards Computational Models of Literary Analysis* 36–43. Genoa, Italy.

Taboada, Maite and Grieve, Jack (2004) Analyzing appraisal automatically. In Y. Qu, J. G. Shanahan and J. Wiebe (eds) *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)* 158–161. Stanford University, CA: AAAI Press.

Taboada, M., Voll, K. and Brooke, J. (2008b) *Extracting Sentiment as a Function of Discourse Structure and Topicality* (Technical Report No. 2008-20): Simon Fraser University.

equinoxonline

Turney, Peter (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, *Proceedings of 40th Meeting of the Association for Computational Linguistics* 417–424. Philadelphia, PA.

Voll, Kimberly and Taboada, M. (2007) Not all words are created equal: Extracting semantic orientation as a function of adjective relevance, *Proceedings of the 20th Australian Joint Conference on Artificial Intelligence* 337–346. Gold Coast, Australia.

White, Peter (2002) Appraisal. In J.-O. Östman and J. Verschueren (eds) *Handbook of Pragmatics* 1–27. Amsterdam: John Benjamins.

White, Peter R. R. (2003) *An Introductory Course in Appraisal Analysis*. Retrieved March 16, 2009 from http://www.grammatics.com/appraisal

Whitelaw, C., Garg, N. and Argamon, S. (2005) Using Appraisal groups for sentiment analysis, *Proceedings of ACM SIGIR Conference on Information and Knowledge Management (CIKM 2005)* 625–631. Bremen, Germany.

Whitelaw, Casey and Patrick, Jon (2004) Selecting systemic features for text classification, *Proceedings Australasian Language Technology Workshop,* 93–100. Sydney, Australia.