Building an annotated dataset of app store reviews with Appraisal features in English and Spanish

Natalia Mora and Julia Lavid
Department of English Studies
Complutense University of Madrid

Abstract

This paper describes the creation and annotation of a dataset consisting of 250 English and Spanish app store reviews from Google's Play Store with Appraisal features. This is one of the most influential linguistic frameworks for the analysis of evaluation and opinion in discourse due to its insightful descriptive features. However, it has not been extensively applied in NLP in spite of its potential for the classification of the subjective content of these reviews. We describe the dataset, the annotation scheme and guidelines, the agreement studies, the annotation results and their impact on the characterisation of this genre.

1 Introduction

Application distribution platforms, or app stores have proliferated in the last decade, allowing users to allow users not only to search, buy, and deploy software apps for mobile devices, but also to share their opinion about the app and other app store products (e.g. films, games, music, et.) in text reviews, not only in English but also in other languages such as Spanish. This is the case of Google's Play Store where app and other product reviews are published online. When users write product reviews, they can either encourage or discourage other users to download the item in question, so these reviews may play a key role in making a product a success or a failure. An example of a typical app review is shown in (1) below:

(1) *Love it... But. I really like this app, it is the best task manager I've had, my phone runs better and I am really maximizing my (limited) storage space. I just wish there would be an ad free version.*

As illustrated by this app review, these texts differ from traditional reviews found in sites like epinions.com in that: a) users have slightly deviated from valuing the items in polarity terms and turned to describing their performance; b) users address directly application' developers; c) users' comments are limited to 1200 characters and, since comments are usually posted via smartphone, typical elements of the internet and mobile language are included, such as abbreviations and emoticons. In addition, sentences frequently miss subjects and links, since authors try to speed up their writing in their phone's small keyboard. All these features make these reviews particularly interesting not only from the linguistic point of view, but also to drive the development effort of app designers and to improve forthcoming releases of a given product.

NLP work on these reviews has mostly focused on extracting patterns related to the length of the review (Vasa et al., 2012) its content (Khalid, 2013), collocation features (Guzman and Maalej 2014), and ambiguity (Islam 2014), and on their polarity on their polarity classification, relying on machine-learning techniques trained over vectors of linguistic feature frequencies (Pang et al., 2002; Finin, 2009), although some more ambitious work has been developed to classify reviews into three and five rating classes using a set of linguistic features including intensification, negation, modality and discourse structure (Brooke 2009). To our knowledge, with the exception of initial work by Taboada and Grieves (2004), there are no studies which explore the potential of Appraisal features to classify and quantify the subjective content of these reviews. This paper, therefore, tries to fill a gap in this area by reporting on the recent

Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pages 16–24
NAACL HLT 2018, New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

development of a bilingual (English-Spanish) dataset of app store reviews annotated with Appraisal tags. We believe that these tags can help categorize the subjective content of these reviews into more fine-grained and diverse features than those focusing only on polarity, quantify the writer's commitment to the opinion, and specify how focused that opinion is.

2 Appraisal

Appraisal is a linguistic theory of subjectivity developed within Systemic-Functional Linguistics to model language's ability to express and negotiate opinions and attitudes within text (Martin 2000; 2003; Martin & White 2005). Appraisal resources are considered as a system of their own within language, and can be divided into three subsystems: *Attitude, Graduation* and *Engagement,* as shown in Figure 1.

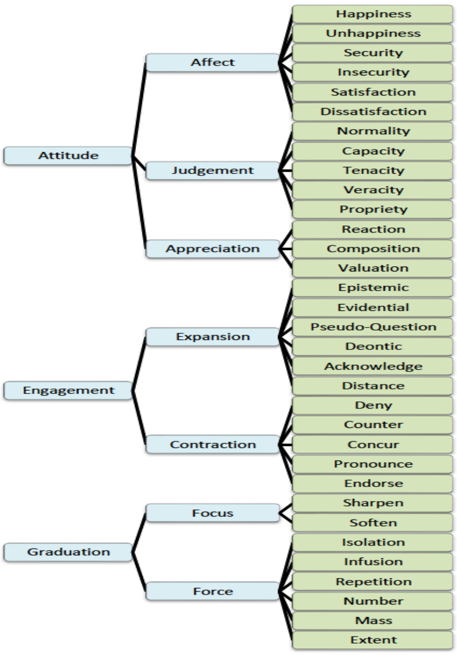


Figure 1: Appraisal subsystems (after Martin and White, 2005)

Attitude is concerned mainly with feelings, such as emotions, judgements and evaluations; it can be further subdivided into *Affect, Judgement* and *Appreciation*, each of which is subdivided into more delicate categories, as shown in Figure 1.

Engagement is concerned with the ways in which the speakers or writers position themselves towards the text and other possible voices, and is further subdivided into *Expansion* and *Contraction,* with more delicate categories expanding them; *Expansion* presents the author's voice as one in a range of possible viewpoints. In *Contraction* the author restricts or challenges other viewpoints; finally, *Graduation* is concerned with the degrees of intensity of the meanings expressed by *Attitude* and *Engagement* realisations, and includes *Focus* and *Force*.

The work developed so far has been mostly circumscribed to Linguistics and basically focused on English, although some cross-linguistic studies involving both European and non-European languages have emerged during the last decade. This includes contrastive work between English and Spanish journalistic texts (Marín and Perucha 2006; McCabe 2007), consumer reviews (Mora 2011, Carretero and Taboada 2009, 2010a, 2010b, 2011, 2014) and other text types (Taboada, Carretero and Hinnel, 2014; Lavid et al. 2014; Lavid, Carretero and Zamorano 2016).

3 Compiling the corpus

In order to compile the bilingual dataset for annotation purposes, the following steps were carried out:

1. A total of 49687 English reviews and 37304 Spanish reviews published before November 2016 were automatically extracted from Google's Play Store using a crawler designed *ad hoc*, as shown in table 1.

	English	**Spanish**
Applications	15721	15225
Games	15288	15328
Books	4909	2223
Films	7793	1595
Music	5976	2933
Total	49687	37304

Table 1: English and Spanish reviews extracted

The reviews included the categories of applications, games, films, books and music. For each category, some of the most famous items were selected (Instagram, Angry Birds, Frozen, Fifty Shades of Grey, Adele, etc.).

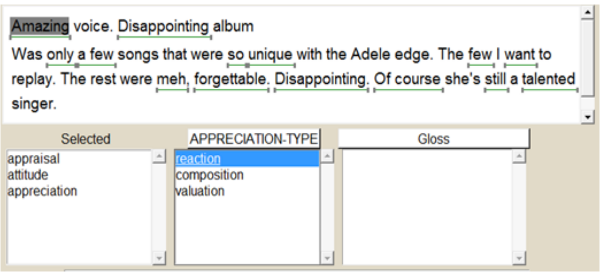
2) From this initial dataset, we randomly selected a smaller set of 250 reviews for annotation purposes, given the amount of effort needed for fine-grained Appraisal annotation. This smaller set contained equal distribution of reviews in terms of language (English and Spanish), similar length, type of polarity (positive or negative) and app category (applications, games, films, books and mu-

sic). When an item had more reviews than needed for the study, those with a higher length were preferred. Thus, the length of the reviews selected ranges from 4 to 240 words, although most of them are about 30-60 words long.

The dataset of 250 reviews was further divided into two smaller sets as follows:

1. An initial training set of 50 reviews was analysed by two annotators. These annotators shared a common background on Spanish and English linguistic studies, both being PhD students in their last year; however, one of them was familiar with the Appraisal Framework while the other one was not. This training set was used to perform agreement studies to validate the annotation scheme and guidelines of Appraisal in English and Spanish.

2. A larger dataset of 200 reviews was annotated by one of those two initial annotators with the Appraisal tags which had been validated through the agreement studies, using the UAM Corpus Tool (O'Donnell 2008) as shown in Figure 2:



Amazing voice. Disappointing album

Was only a few songs that were so unique with the Adele edge. The few I want to replay. The rest were meh, forgettable. Disappointing. Of course she's still a talented singer.

Figure 2: Annotation interface

4 Annotation methodology

We applied the annotation steps suggested by Lavid (2017, 2012) and Hovy and Lavid (2010), as follows:

a) An annotation scheme and guidelines were designed on the basis of the main features proposed in *Appraisal* Theory, along the three axes of *Attitude*, *Engagement* and *Graduation*. This is described in 4.1.

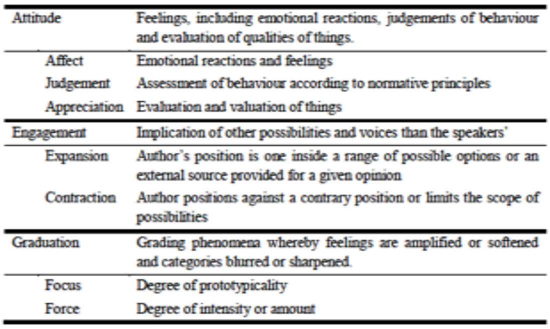
b) Agreement studies were designed to test the empirical validity of the annotation scheme. These were carried out by two independent annotators working separately on a training corpus of fifty mobile application reviews. This is described in 4.2. and 4.3.

c) On the basis of the results of the agreement studies, a larger corpus of two hundred reviews was single-annotated with the validated *Appraisal* tags of the annotation scheme. The results of this annotation is described in 4.4.

d) The distribution of Appraisal tags was examined in the English and the Spanish reviews in order to obtain a characterisation of this genre. This is described in 4.5.

4.1 Annotation scheme and guidelines

On the basis of the Appraisal tags proposed by Martin and White (2005), we designed an initial annotation scheme, consisting of a more general core tagset, and an extended tagset, with some more delicate features. The core tagset was common to English and Spanish and is presented in table 2.



Attitude	Feelings, including emotional reactions, judgements of behaviour and evaluation of qualities of things.
Affect	Emotional reactions and feelings
Judgement	Assessment of behaviour according to normative principles
Appreciation	Evaluation and valuation of things
Engagement	Implication of other possibilities and voices than the speakers'
Expansion	Author's position is one inside a range of possible options or an external source provided for a given opinion
Contraction	Author positions against a contrary position or limits the scope of possibilities
Graduation	Grading phenomena whereby feelings are amplified or softened and categories blurred or sharpened.
Focus	Degree of prototypicality
Force	Degree of intensity or amount

Table 2: Core tagset of annotation schema

4.2 Agreement studies

Three experiments (also called 'agreement studies') were designed to test the reproducibility of the scheme's tags. The first experiment focused on the identification of the spans or markables, the second one addressed the selection of the three main general types of *Appraisal*, and in the third one, coders had to make fine-grained selections from the more delicate subtypes.

The purpose of the first experiment was to investigate which elements were considered as *Appraisal* tags by two coders working independently and to delimit their boundaries. Here coders were instructed to annotate the shortest lexical span expressing *Appraisal*, although one of them was familiar with the theory before the experiment.

Once coders agreed on the spans, the second annotation experiment addressed the labelling of the *Appraisal* markables with one of the three coarser tags and their main subtypes, i.e.: *Attitude* (*Affect, Judgement and Appreciation*), *Engagement* (*Expansion and Contraction*) or *Graduation* (*Force and Focus*).The purpose of this experiment was to investigate whether coders could distinguish among the different coarse tags and their subtypes, before getting deeper into more delicate categories. If significant inconsistencies were found, this step would make it easier to identify any conflictive or confusing aspects of the theory or the guidelines.

In the third annotation experiment, coders were instructed to use more fine-grained tags from the extended tagset to label the selected markable. These include tags such as *Happiness, Unhappiness, Security, Insecurity, Satisfaction* and *Dissatisfaction* in the case of *Affect*; *Normality, Capacity, Tenacity, Veracity* and *Propriety* in the case of *Judgement*; *Reaction, Composition* or *Valuation* in the case of *Appreciation*; *Epistemic, Evidential, Pseudo-Question, Deontic, Acknowledge* and *Distance* in the case of *Expansion; Deny, Counter, Concur, Pronounce* and *Endorse* in the case of *Contraction*; *Sharpen* and *Softer* in the case of *Focus*; and *Isolation, Infusion, Repetition, Number, Mass* and *Extent* in the case of *Force*. The purpose of this experiment was to investigate whether highly delicate categories could be coded consistently by two independent coders, and whether subtle differences in meaning could be distinguished.

4.3 Results of agreement studies

The results of the first experiment yielded a substantially high degree of agreement between coders (Kappa=0.86), although some disagreements also occurred in a small percentage of the cases (4%). These cases occurred when the span was either selected by one of the coders and not by the other, or when the span's length was different. Most of the cases of disagreement occurred in long and complex sentences that do not directly reflect an opinion, but must be contextualised to convey an evaluative meaning, as in (2) below:

(2) *Vale la fama que tiene* [translation: it's worth its popularity] (T43): In this example one coder selected the full phrase while the other one selected only the verb 'vale' [it's worth it].

In the second agreement study the agreement between coders was even higher (Kappa= 0.96). The increase in the k-value was probably due to the fact that the span selection was already decided. Although coders could in most cases distinguish between the three major categories of *Attitude*, *Engagement* and *Graduation*, the highest mismatches were found when coding *Graduation* followed by *Engagement* and *Attitude*.

Graduation appeared as the most conflictive category, which points to an unclear difference between intensification or additional description and values.

As to disagreements found between different subtypes of categories (i.e.: *Affect, Judgement, Appreciation, Expansion, Contraction, Force* and *Focus*), the category with most conflictive cases was *Attitude: Appreciation*, which was mostly confused with other subtypes of *Attitude*. The second highest disagreements were found within the category of *Graduation*, with more cases confusing *Force* with other subtypes than *Focus*, followed closely by *Engagement*, where *Contraction* was more often confused with other categories than *Expansion*. Figure 3 graphically displays the distribution of these disagreements.

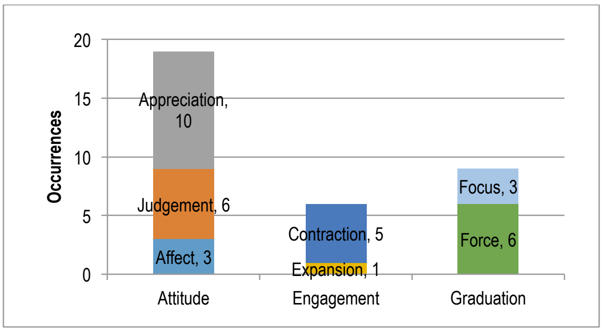
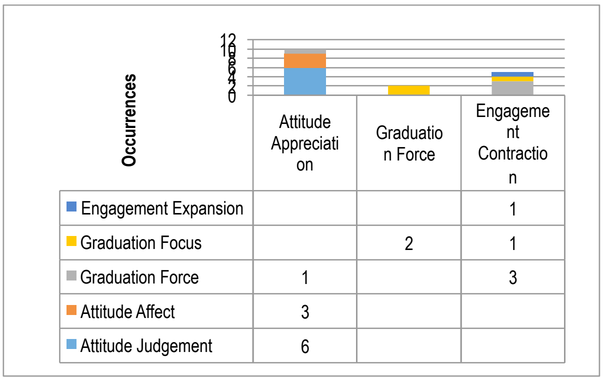


Figure 3: Disagreements among Appraisal subtypes in the bilingual corpus

Finally, Figure 4 shows the distribution of the most controversial combinations, that is, which ones were typically used one instead of the other.



	Attitude Appreciati on	Graduatio n Force	Engageme nt Contractio n
🟦 Engagement Expansion			1
🟨 Graduation Focus		2	1
▨ Graduation Force	1		3
🟧 Attitude Affect	3		
🟦 Attitude Judgement	6		

Figure 4: Disagreements in combination of categories

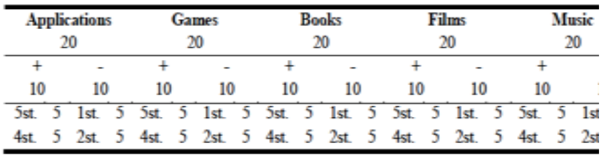
The combinations which caused more disagreement were *Attitude-Appreciation* and *Attitude-Judgement,* since they were often confused by coders. Theoretically, *Judgement* refers to other people's behaviour while *Appreciation* focuses on objects and natural phenomena. However, evaluative elements on moral aspects, typically used for human beings, can be associated with objects in a metaphorical way. Examples which caused disagreement were the use of adjectives such as '*flojísima*' [transl. 'very poor'], '*lenta'* [slow], referring to a novel; '*kid-friendly'* or ˊ*sweet* referring to a film. *Attitude-Appreciation* was also confused with *Attitude-Affec*t and vice versa in several cases, probably due to the fact that it is not clear when the focus is on the object causing a feeling or the author having that feeling caused by the object. An example would be the use of *'stunned'.*

The tags of *Graduation-Force* and *Engagement-Contraction* also caused disagreement between coders, as in the case of the item '*really*', which has different meanings that are not always clearly distinguishable.

In the third agreement study, the agreement was only moderate (kappa=0.49). Most disagreements were caused by the difficulty to discriminate among the different subtypes of *Attitude*. The categories which caused more disagreement were *Reaction* and *Valuation*, which were coded differently on several occasions. Thus, for example, in the case of adjectives such as '*pobre'* [poor], or '*lovable*', coders hesitated between considering them as qualities of the object (valuation) to which they were assigned, or a consequence of the user's feelings (reaction).

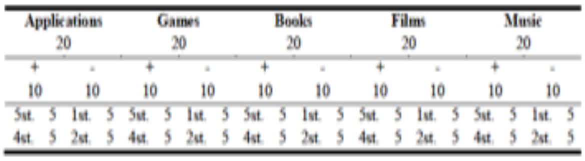
4.4 Annotation of the larger dataset

Our next step was to annotate a larger dataset with the validated tags of the proposed annotation scheme. This consisted of two hundred texts filtered and selected following the same procedure as the training set: it included comparable English and Spanish texts evenly distributed, as illustrated in Tables 3 and 4 (st. stands for 'stars', regarding the 1-to-5 star rating):



Applications				Games				Books				Films				Music			
20				20				20				20				20			
+		-		+		-		+		-		+		-		+			
10		10		10		10		10		10		10		10		10			
5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st	
4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st	

Table 3: English dataset



Applications				Games				Books				Films				Music			
20				20				20				20				20			
+		-		+		-		+		-		+		-		+		-	
10		10		10		10		10		10		10		10		10		10	
5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st.	5	5st.	5	1st.	5
4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st.	5	4st.	5	2st.	5

Table 4: Spanish dataset

The texts addressed several items inside each of the products in order to enhance diversity in the texts. The reviews addressed at least two items per category, including applications such as *Clean Master, Instagram*, games such as *Angry Birds, Candy Crush,* books such as *All the Light We Cannot See, Fifty Shades of Grey, The Girl on the Train,* films such as *Avatar, Gravity, Frozen, The Wolf of Wall Street,* and music such as AC/DC or Adele.

The annotation tool was the UAM Corpus tool5, a free state-of-the-art annotation platform which supports annotation of multiple texts at multiple linguistic levels (clause, sentence, document, etc.) as well as analysis methods such as instances retrieval and statistical measurements. Also, the first author of this paper single-annotated this larger set, instead of double annotating and adjudicating, following Dligach et al.'s (2010) suggestion, according to which "it is often better to single annotate more data because it is a more cost-effective way to achieve a higher performance".

4.5 Annotation results

At a general level, the most frequently annotated category was *Attitude* (40.89%), followed by *Engagement* (35.64%) and *Graduation* (23.46%). However, when looking at the more specific tags, the most frequent one was *Contraction* (26.93%). This is due to the number of negations (*Deny*) (3a, 3b) and hypothetical situations (*Counter*) (4a, 4b) that are included in both languages. The second most common category was *Appreciation* (24.59%) (5a, 5b), which should be expected since the annotated texts are rich in *Valuation* or expressions conveying a value associated with an object and their aim is to describe those reviewed items. Finally, the tag with the third highest number of occurrences was *Force* (23.14%) (6a, 6b), which includes all those intensifiers and quantifiers that increase or lower the value of other nouns, adjectives or verbs.

(3a) At the beginning, neither is believable (79)

(3b) *A mi no me da ningun problema* (128) [translation: it doesn't give me any problem]

(4a) However, this is … (74)

(4b) *Pero realmente lo unico que quieren*…(148) [translation: but what they only really want…]
(5a) This is amazing (84)
(5b) *Muy sobrevalorada* (168) [translation: very overrated]
(6a) Shame after so long a wait (90)
(6b) *Muy cara* (200) [translation: very expensive]

The category of *Focus* (0.32%) showed a very low distribution, probably because it is used to soften or sharpen the boundaries of a word, i.e., to express how close it is to the prototypical idea of that item, but users prefer to quantify nouns rather than stress or diminish their core meanings. *Judgement* (6.19%) is used to assign social or moral values to people but here it was used not only to address people but also objects. In any case, this kind of value was not a pivotal one in the items selected. Thirdly, expressions of *Expansion* (8.72%), showing different levels of certainty and allowing for other opinions apart from the autorial one, only appeared in half of the occasions in comparison with *Contraction*. This means that reviewers place the stress on their own voice, limiting the possibilities of other options, instead of presenting their opinion as one of a range of possible choices. Finally, *Affect* is placed in the very middle of the ranking (10.11%). This type of expressions refers to someone's feelings, how the author (or other users) feel with respect to the item reviewed and, in spite of their occurrence in the annotated texts, reviews focus much more on the value or even the effects of the item itself than on users' feelings.

The distribution of these categories in the larger corpus is graphically displayed in Figure 5:

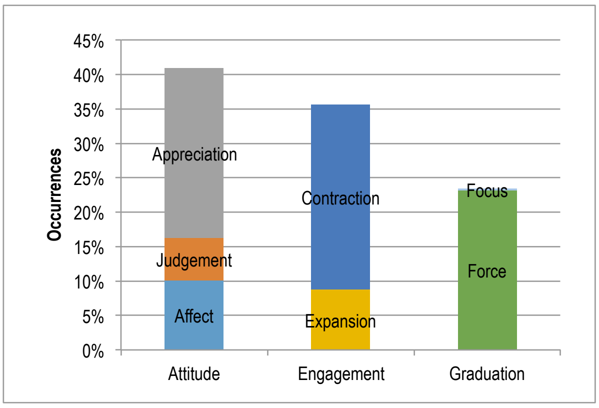


Figure 5: Distribution of Appraisal tags in the larger corpus

As to the language-specific preferences, English shows a slightly higher preference for *Engagement* (36.92%) than Spanish (33.94%), as well as for *Graduation* (24.69% in English vs. 21.82% in Spanish), although these are not statistically significant. However, the most visible difference involves *Attitude*, where it was found that Spanish occurrences go up to 44.24% while only a 38.39% of the English tags are marked as *Attitude*.
When comparing the preferences in the use of specific *Appraisal* tags in the different products (applications, games, books, films, music, etc.), the initial results presented a small difference between two groups: the group of applications and games were characterised by a higher use of *Engagement* resources, while the group of books, films and music showed a more frequent use of *Attitude* categories. *Graduation* did not present a clear tendency with similar results in both groups except for a wider use in games. *Appreciation* was ranked first as an *Attitude* resource in films and books, while *Affect* occurs more frequently in games, as well as in applications and books. *Judgement* is more likely to appear in applications than in any other item reviewed.

Reviews can also be classified according to the number of stars they assign to the item reviewed. This rating goes from 1 to 5 stars, Our corpus was divided in two groups: one group of one hundred negative reviews (with 1 and 2 stars) and another group of one hundred positive reviews (with 4 and 5 stars). While *Attitude* resources were more abundant in positive reviews, negative reviews use *Engagement* expressions more frequently. The use of *Graduation*, on the other hand, is virtually the same in both groups. Also, although the distribution of *Affect* is almost identical in positive and negative reviews (24.07% vs. 25.47%), positive reviews abound in *Appreciation* (64.93% vs. 54.67%), while negative ones make a wider use of *Judgement* (11.00% vs. 19.86%). This is related to the specific use that authors make of *Judgement* expressions, focusing on *Propriety* and, more specifically, negative spans. By contrast, when users are giving a positive evaluation of the item, they use more *Appreciation* resources to describe and justify the positive rating.

5 Summary and Discussion

The results of the annotations in the larger set indicate interesting tendencies in the distribution of *Appraisal* tags in the English and Spanish reviews, although they were not statistically significant. First, mobile applications reviews were shown to be especially rich in *Attitude* tags, followed by *Engagement*, while *Graduation* tags occur much less frequently. This distribution reflects the communicative purpose of these texts, which is to present users' opinions on a given product. Therefore, the majority of the *Appraisal* tags are expressions of *Attitude* which assign a value to the item reviewed, or express someone's feelings related to that item. The need to engage other users in the reviews is also reflected in the quite abundant use of *Engagement* tags in both the English and the Spanish reviews. *Graduation* tags, used to intensify or soften ideas, appear much less frequently in these reviews, indicating that users prefer other *Appraisal* strategies to convey their opinions on a given product.

As to the preferred tags from the extended tagsets, the reviews are rich in *Appreciation* expressions as they focus on the product, including its performance, qualities, effects, etc., while expressions of *Affect* and *Judgement* are less frequently used comparatively. The most frequent subtypes of *Appreciation* tags are *Valuation* and *Reaction*, while *Composition* (how the object is composed) is less frequently used. *Affect* tags are also common, but not as much as *Appreciation*. *Affect* deals with feelings and emotions, expressing the way the author or someone else feels in relation to the product reviewed and are the second-most common subtype of *Attitude* markers in the bilingual corpus. Their role in the reviews is usually supportive with respect to the role of *Appreciation* tags: if the qualities of the object itself are not enough to show why someone's opinion is the way it is, the expression of the users' feelings supports the emotional aspects of their opinion. The most common subtypes of *Affect* used in the larger corpus are *(Un)Happiness* and *(Dis)Satisfaction*. These include messages about how much users like (or dislike) the product or how satisfied and interested they are. Usually, authors tend to include *Happiness* expressions more often than *Unhappiness* elements, although *Satisfaction* and *Dissatisfaction* do not show such a clear distinction. (*In)Security* messages are not recurrent in these texts, so meanings related to fear, surprise, trust and the like are not frequently assigned to these products.

Judgement is the least used category in *Attitude*, probably because it includes meanings used to evaluate people's behaviour and not objects or products. Despite this fact, more occurrences have been found than expected, as when users focused on meanings related to *Capacity* and *Propriety*, classifying a game's bugs as a '*theft*', a character's behaviour as '*reprehensible*' or a singer as '*(un)talented'*.

Engagement, as mentioned before, had the second highest rate after *Attitude*, and it is divided into two main categories: *Expansion,* which presents the author's voice as one in a range of possible voices, and *Contraction*, which delimits and denies other possible voices. *Expansion* was the least frequent choice, while *Contraction* types are highly frequent in the reviews mainly due to *Disclaim* elements (*Counter* and *Deny*), which include common linguistic items such as conjunctions and negative particles.

Finally, *Graduation* was the least used category, with *Force* tags outnumbering *Focus* ones. This is probably because these reviews do not usually modulate the level of prototypicality of the nouns they use to name entities, but they intensify adjectives, verbs and indicate quantities for nouns. Thus, a product is not good but *very* good, a bug did not just happen, but happened *many* times, and they do not just like it, but like it *a lot.*

With respect to the language-specific comparisons, the Spanish reviews use *Attitude* resources much more frequently than the English ones, which prefers *Engagement* and *Graduation* elements. Thus, while the Spanish reviews draw on feelings and qualities, the English ones modulate their voice inside the text through *Engagement* as well as through expressions of *Graduation*. Expressions of *Satisfaction* were more frequently used in the Spanish reviews, while the English ones focused on those expressing *Happiness*. English writers frequently used words like '*love'* and '*like'* for any kind of product while Spanish writers use '*agradecer'* [thank] or '*esperar'* [hope]. Similarly, Spanish writers have a higher interest in describing *Capacity* and *Valuation*, whereas English ones lean more strongly on expressions of *Normality, Veracity* and *Reaction*.

As to the distribution of *Engagement* features, English reviews modulate certainty more extensively through *Epistemic* tags and *Pseudo-Questions* and are also more sarcastic by using rhetoric questions in their writings, while Spanish ones are much more direct using *Deontic* resources and basing their opinion on empirical sources. Along the same lines, *Counter* elements are more profuse and varied in English, showing opposition and contrast, while Spanish writers are more direct by simply rejecting any other possibilities by means of *Deny* resources.

With respect to differences among products, the observed distributions allowed the grouping of some products: one formed by applications and games with a higher use of positive *Affect* categories, such as *Happiness, Security* and *Satisfaction*; and a second one formed by books and films, which abound in negative ones such as *Unhappiness* and *Insecurity.* Music shares some characteristics with both groups but it has its own proper qualities.

With respect to the differences between positive and negative reviews, negative reviews abound in expressions of *Judgement* that is not observed in positive texts. As mentioned above, *Judgement* expressions typically address morally incorrect behaviours (*Propriety*), since positive moral actions are taken for granted. Positive *Judgement* realisations usually address *Capacity* meanings, such as talent, an adequate operation or improvements made in a product. The most common Affect meanings are *Happiness* and *Satisfaction* in positive reviews and, unsurprisingly, *Dissatisfaction* and *Unhappiness* in the negative ones.

Positive reviews present a higher use of *Epistemic* and *Deontic* resources with authors introducing their opinions by means of spans like *'I think'* and also recommend the product to other users through obligation meanings like *have to*. Negative reviews use *Pseudo-Questions* and *Evidential* markers since they distance from the item by means of sarcastic questions or use verbs like '*seem'* to introduce a negative quality instead of stating it directly. *Counter* realisations were far more common than the other *Disclaim* type, *Deny*, in positive reviews, but they both presented similar percentages in negative reviews. This is due to a higher use of negative elements in negative reviews, as can be expected, instead of a much lower use of *Counter* items.

Finally, *Graduation* differences include a higher use of *Isolation* modifiers in positive reviews, and a more profuse use of *Number* items in negative ones. This means that words like '*so'* or '*very'* are typically attached to positive expressions like '*good'* instead of '*bad'*, while '*many'*, '*some'*, etc. are used when criticising a product.

6 Concluding remarks

The work reported in this paper on the annotation of a bilingual (English-Spanish) dataset of mobile application reviews with *Appraisal* features has shed light on a number of theoretical and applied issues which deserve research attention in the Natural Language Processing (NLP) and the Linguistics communities. From the theoretical perspective, the empirical validation of the annotation scheme will contribute to the refinement and reformulation of certain *Appraisal* features which have proved problematic in the annotation of the genre of mobile application reviews; and it will hopefully encourage further applied work to other genres and other languages. From the applied NLP perspective, the creation of a bilingual (English-Spanish) dataset containing *Appraisal* features will hopefully be useful for the development of machine learning algorithms for large scale annotation of this genre and other possible ones in the near future.

Future work will be focused on investigating the realisation of *Appraisal* in long phrases and sentences, in order to find common validated features beyond readers' interpretations. Another interesting line of future research is the extension of the empirical validation of more delicate *Appraisal* features for which insufficient evidence was found in the current corpus. It would also be relevant to extend the current range of items reviewed to a wider range of products in order to find possible groupings that share *Appraisal* features, thus confirming or diverging from the tendencies pointed out in this work.

References

Bloom, K., N. Garg and S. Argamon. 2007. Extracting Appraisal Expressions. In *Proceedings of NAACL HLT*. Rochester. 308–315

Bloom, K., S. Stein and S. Argamon. 2007. Appraisal Extraction for News Opinion Analysis at NTCIR-6. In *Proceedings of NTCIR-6 Workshop Meeting*. Tokyo, Japan.

Brooke, Julian. 2009. A Semantic Approach to Automatic Text Sentiment Analysis. M.A. thesis, Simon Fraser University, Burnaby, B.C., Canada.

Carretero, M. and M. Taboada. 2009. Contrastive analyses of Evaluation in text: Key issues raised by the application of Appraisal Theory to corpora of consumer-generated product reviews in English and Spanish. *I CongresoInternacional de Lingüística de Corpus.* Murcia, Spain.

Carretero, M. and M. Taboada. 2010a. Products, consumers and evaluation: a proposal of solutions to problematic issues of Attitude in English and Spanish consumer reviews. *22nd European Systemic Functional Linguistics Conference and Workshop.* University of Primorska, Slovenia.

Carretero, M. and M. Taboada. 2010b. The annotation of Appraisal: How Attitude and epistemic modality overlap. *4th International Conference on*

Modality in English (ModE4). Madrid, Spain.

Carretero, M. and M. Taboada. 2011. Annotating Appraisal: contrastive issues raised in the analysis of consumer reviews in English and Spanish. *38th International Systemic Functional Linguistic Congress.* Lisbon, Portugal.

Dligach, D., R.D. Nielsen and M. Palmer. 2010. To annotate more acccurately or to annotate more. In *Proceedings of the Fourth Linguistic Annotation Workshop.*64{72}, 55, 66, 132, 337, 380, 389.

Hovy, E., and Lavid, J. 2010. Towards a 'Science' of Corpus Annotation: A new Methodological Challenge for Corpus Linguistics. *International Journal of Translation,* 22 (1). 13-36.

Lavid, J. 2012. Corpus analysis and annotation in CONTRANOT: Linguistic and Methodological Challenges. In I. Moskowich and B. Crespo (eds.), *Encoding the past, decoding the future: corpora in the 21st century*. Cambridge: Cambridge Scholars. 205-220.

Lavid, J. 2017. Annotating complex linguistic features in bilingual corpora: The case of MULTINOT. In T. Declerck and S. Kübler (eds.), *Proceedings of the Workshop on Corpora in the Digital Humanities (CDH 2017)*. IN, USA: Bloomington. Available online athttp://ceur-ws.org/Vol-1786.

Lavid J., Carretero, M. and JR. Zamorano (2016): Contrastive Annotation of Epistemicity in the Multinot Project: Preliminary Steps. In Harry Bunt (ed.). *Proceedings of the ISA-12, Twelfth Joint ACL - ISO Workshop on Interoperable Semantic Annotation,* held in conjunction with *Language Resources and Evaluation Conference 2016*. 81-88.

Marín-Arrese, J.I. and B. Núñez Perucha. 2006. Evaluation and Engagement in Journalistic Commentary and News Reportage. *Revista Alicantina de Estudios Ingleses*19. 225-248.

Martin, J. and White, P. 2005. *The language of evaluation: Appraisal in English*. Palgrave, Macmillan.

Mora, N. 2011.*Annotating Expressions of Engagement in Online Book Reviews: A Contrastive (English-Spanish) Corpus Study for Computational Processing*. http://eprints.ucm.es/13754.

O'Donnell, M. 2008. "The UAM CorpusTool: Software for corpus annotation and exploration". Proceedings of the XXVI Congreso de AESLA, Almeria, Spain, 3-5 April 2008.

Pang, B., L. Lee and S. 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, EEUU.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2). 1–135.

Taboada, M. and J. Grieve. 2004. Analyzing appraisal automatically. In *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07).*). Stanford University, CA, EEUU. 158-16.

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

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

Whitelaw, C., Garg, N. and S. Argamon. 2005. Using appraisal taxonomies for sentiment analysis. In *Proceedings of MCLC-05, the 2nd Midwest Computational Linguistic Colloquium*, Colombia, EEUU.