Lang Resources & Evaluation
DOI 10.1007/s10579-011-9140-5

ORIGINAL PAPER

EmoTales: creating a corpus of folk tales with emotional annotations

Virginia Francisco · Raquel Hervás · Federico Peinado · Pablo Gervás

Abstract Emotions are inherent to any human activity, including human–computer interactions, and that is the reason why recognizing emotions expressed in natural language is becoming a key feature for the design of more natural user interfaces. In order to obtain useful corpora for this purpose, the manual classification of texts according to their emotional content has been the technique most commonly used by the research community. The use of corpora is widespread in Natural Language Processing, and the existing corpora annotated with emotions support the development, training and evaluation of systems using this type of data. In this paper we present the development of an annotated corpus oriented to the narrative domain, called EmoTales, which uses two different approaches to represent emotional states: emotional categories and emotional dimensions. The corpus consists of a collection of 1,389 English sentences from 18 different folk tales, annotated by 36 different people. Our model of the corpus development process includes a post-processing stage performed after the annotation of the corpus, in which a reference value for each sentence was chosen by taking into account the tags assigned by annotators and some general knowledge about emotions, which is codified in an ontology. The whole process is presented in detail, and revels significant results regarding the corpus such as inter-annotator agreement, while discussing topics such as how human annotators deal with emotional content when

V. Francisco (✉) · R. Hervás · F. Peinado · P. Gervás
Departamento de Ingeniería del Software e Inteligencia Artificial, Facultad de Informática, Universidad Complutense de Madrid, Madrid, Spain
e-mail: virginia@fdi.ucm.es

R. Hervás
e-mail: raquelhb@fdi.ucm.es

F. Peinado
e-mail: email@federicopeinado.com

P. Gervás
e-mail: pgervas@sip.ucm.es

Published online: 23 February 2011

performing their work, and presenting some ideas for the application of this corpus that may inspire the research community to develop new ways to annotate corpora using a large set of emotional tags.

Keywords Text corpora · Corpus annotation · Emotional ontology · Emotional categories · Emotional dimensions

1 Introduction

Emotions are inherent to any human activity, including our interactions with computers. Identifying emotions expressed in natural language within a speech or text document is becoming a required feature for any computational system that aims to offer a more natural interface to its users. For example, recognizing certain emotions in a human speaker would permit a computer to react to her commands according to her personal situation, instead of giving a neutral response; and that response from the machine can also be modified after taking into account the emotion it should express (Evens 2002; Krenn et al. 2002). Synthesized speech would also be significantly improved by reproducing different emotional connotations when modulating the synthesized voice.

The recognition of emotions expressed in natural language is not only important for classical interfaces but also for on-line advice and recommendation systems (Pang and Lee 2008). The interest that users show in on-line opinions and the potential influence of such opinions is something that vendors are paying more and more attention to Hoffman (2008), making it very important to identify the emotions behind them automatically. Emotional analysis algorithms have been recently applied to the creation of computational models of human opinion from customers' on-line reviews (Wright 2009), for example.

In addition, the automatic generation of text and speech has been widely developed over the last two decades, usually giving rise to technological solutions for restricted domains. Affective Computing aims for more natural interactions, particularly in the areas of the recognition and generation of emotions (Pang et al. 2002; Turney and Littman 2003; Campbell 2005; Wiebe et al. 2005; Merota 2007). However, despite the current interest in emotion tagging in the last decade, the number of existing resources is quite poor. The corpus presented in this paper aims to be a useful contribution towards these needs, and starts by from studying the problems associated with the annotation of emotions.

Annotating text with emotional content is a difficult task. As the identification and assignment of emotions are subjective decisions, it is common that different human annotators assign different emotional tags to the same sentence or piece of text. Therefore, it is very important to study how the emotional annotation process of a corpus is performed in order to define that process properly, while reducing its dependency on subjective criteria as much as possible. In such a context, it would also be extremely useful if both the representation chosen and the annotated material were devised in such a way as to provide flexible transitions between different degrees of granularity in the annotation. Because some research efforts

may wish to concentrate on a small set of basic emotions, and others may want to consider a broader range, a resource that allows easy conversion from annotations in terms of basic emotions to annotations in terms of larger sets of emotional labels, or conversion across different methods of representing emotion, would be very useful.

Our research goal is therefore to create an annotated corpus for narrative applications using two different and relevant approaches to represent emotional states: *emotional categories* and *emotional dimensions*. For the annotation of emotional content, we combine different approaches for representing emotions and use an ontology as a knowledge-intensive resource. According to the classification of language resources presented in Witt et al. (2009), the corpus we are presenting is a *text-based static resource*, because emotions are annotated using the original texts, which are presented sentence-by-sentence to the annotators; the result acts as an inventory of data instead of a processing or editing tool.

The rest of the paper is structured as follows. Section 2 presents a review of the terminology used in the field in which our work is included, a brief review of definitions of emotions, different methods used to structure them, and some related corpora that have been described in the scientific literature. Sections 3 and 4 explain respectively the main ideas taken into account during the design of EmoTales, our corpus, and how the annotation process was performed. Section 5 presents the various post-processing steps required to obtain an emotional reference value based on the annotations for each sentence in the corpus. In Sect. 6 we evaluate the annotation of the corpus by considering inter-annotation agreement for both emotional categories and dimensions. Section 7 discusses the contributions of the approach proposed. Finally, in Sect. 8 we conclude with contributions to the methodology of creating text corpora annotated with emotions, and draft some lines of future work that have to be considered in order to apply and improve our results.

2 Related work

There has been to date no uniform terminology established for the relatively young field in which our work is included. The work that deals with the computational treatment of *opinion, sentiment* and *subjectivity* in text has come to be known as *Opinion Mining, Sentiment Analysis* and/or *Subjectivity Analysis* (Pang and Lee 2008). The terms *Review Mining* and *Appraisal Extraction* have been also used, and there are some connections to *Affective Computing*, where goals include enabling computers to recognize and express emotions (Picard 1997). This proliferation of terms reflects differences in the connotations that these terms carry.

The term *Opinion Mining* first appears in Dave et al. (2003). According to the authors, the ideal opinion-mining tool would "process a set of search results for a given item, generating a list of product attributes and aggregating opinions about each of them".

Sentiment Analysis is considered the computational treatment of opinion, sentiment and subjectivity in text. A fundamental technique in *Sentiment Analysis* is the *classification of emotions* (Pang and Lee 2008), usually by asking questions such as "what emotion is evoked by this sentence?". The manual classification of

this kind of information typically leads to useful corpora for the research community.

Affect Sensing is the ability of a computer to remain aware of its user's affective states and transitions (Barreto 2008), which is essential to any kind of affective computing. In fact, Picard (2003) has identified sensing and recognizing emotion as one of the key challenges that must be conquered to bring the full promise of affective computing to fruition. Dictionaries of affect offer great promise to lexical affect sensing since they contain information on the affective qualities of single words that may be employed to estimate the emotional tone of the corresponding text. On the other hand, *Emotion Detection* (Barreto 2008) is the ability of a computer to remain aware of its user's emotions.

As we can see, there is no clear distinction between the different fields related with emotions. For example, when broad interpretations are applied, *Sentiment Analysis* and *Opinion Mining* can be considered the same field of study.

In the following subsections we discuss the definition and formal representation of emotions. We also review the most relevant text corpora with emotional markup that can be found in the literature for further comparison to our work.

2.1 Definitions of emotion

There are plenty of emotional models proposed in the academic literature (Scherer 1984; Cowie et al. 1999; Parrott 2001; Cowie and Cornelius 2003), but there seems to be agreement about the fact that emotions are subjective experiences each person experiences in a very different way. Following this idea, Parrott (2001) defines emotion as a response to events that are important to us. These responses are governed by specific laws and they emerge and manifest themselves according to what the operating mechanism of these laws dictates. In the context of this paper the emotional states defined by Cowie and Cornelius (2003) are the type of emotional content considered.

In order to study emotional states we must decide how they are going to be represented. There are different ways to represent them (Cowie and Cornelius 2003), but two representation techniques are the most commonly accepted: *emotional categories* and *emotional dimensions*.

Emotional categories are the most usual method for describing emotions. This method is based on the use of emotional tags. Different languages provide assorted words with varying degrees of expressiveness for the description of emotional states. That is why several approaches have been proposed to reduce the number of words used to identify emotions, for example with the use of *basic emotions, super ordinate emotional categories* or *essential everyday emotion terms*. *Basic emotions* refer to those that are more well-known and understandable for everybody than others (Cowie and Cornelius 2003). In the *super ordinate emotional categories* approach some emotional categories are proposed as more fundamental, with the argument that they subsume the others (Scherer 1984). Finally, the *essential everyday emotion terms* approach focuses on emotional words that play an important role in everyday life (Cowie et al. 1999).

Emotional dimensions are measures that try to model the essential aspects of emotions numerically. Although there are different dimensional models with different dimensions and numerical scales (Fontaine et al. 2007), most of them agree on three basic dimensions called *evaluation, activation* and *power* (Osgood et al. 1957). *Evaluation* represents how positive or negative an emotion is. At one extreme we have emotions such as *happiness, satisfaction* and *hope* while at the other we find emotions such as *unhappiness, dissatisfaction* and *despair*. *Activation* represents an activity versus passivity scale of emotions, with emotions such as *excitation* at one extreme, and at the other emotions such as *calmness* and *relaxation*. *Power* represents the sense of control which the emotion exerts on the subject. At one end of the scale we have emotions characterized as completely controlled, such as *fear* and *submission* and at the other end we find emotions such as *dominance* and *contempt*.

The clearest distinction between the two methods is that emotional dimensions describe a continuous space as opposed to the discrete space that is described by emotional categories.

2.2 Structuring the space of emotions

Psychologists have been searching for a suitable way to structure our emotional repertoire. Several methods have been proposed, each with its own advantages and disadvantages.

Methods based on emotional dimensions aim to capture the similarities and differences among emotions. Some researchers propose a two-dimensional space that exclusively considers the emotions of evaluation and activation. This is called the *circumflex model* where the points that correspond to all possible emotions form a circle (Russell 1980). Viewing the multitude of emotions as points in a two-dimensional space can be useful in understanding the most generic emotions but not the most specific ones. This model reduces the variety of emotional states, and does not capture the slight differences found beyond the most generic sensations.

As an alternative to dimensional spaces some researchers have used cluster analysis (Storm and Storm 1987; Shaver et al. 1987; Parrott 2001; Aristotle 1960). These approaches group emotions into clusters, with the number of clusters depending on each specific approach. Storm and Storm (1987) proposes the use of 12 clusters: *love, happiness, sadness, anger, fear, anxiety, contentment, disgust, hostility, liking, pride* and *shame*. Shaver et al. (1987) proposes the use of 5 clusters called *affection, happiness, sadness, anger* and *fear*. Parrott (2001) presents a more detailed list of emotions categorized in a short tree structure. This structure has three levels for primary, secondary and tertiary emotions. As primary emotions, Parrot presents *love, joy, surprise, anger, sadness* and *fear*. Secondary emotions give nuance to primary emotions, e.g. *love* has *affection, lust* and *longing* as secondary emotions. Finally, tertiary emotions give further nuance to secondary emotions, e.g. *lust* is a secondary emotion with *arousal, desire, passion* and *infatuation* as tertiary emotions. Aristotle (1960) uses 11 basic emotions: *anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love* and *sadness*.

Instead of grouping emotions according to their global similarity, other researchers prefer to group emotions based on different criteria such as the components of their appraisals (Scherer 1984) or the events that give rise to them (Ortony et al. 1988).

To summarize, there are many different ways to structure emotions and each approach may be useful for a different purpose. Any approach that aims to be useful in a great variety of applications should take advantage of all these different representations of the world of emotions.

2.3 Emotional text corpora

According to Douglas-Cowie et al. (2003), when studying text corpora annotated with emotions, it is important to consider three main aspects: the scope of the corpus (i.e. the emotional classes under study, the number of annotators and the language of the documents), the context of the resource (i.e. in-isolation or in-context annotation) and the descriptors used for the annotation (i.e. emotional categories or emotional dimensions). In Table 1 we show some representative corpora in this field, and provide information about these three aspects. A more thorough review can be found in Pang and Lee (2008).

With respect to the emotional descriptors employed, most of these corpora are mainly oriented towards the annotation of the evaluation dimension, and those dealing with emotional categories consider a small list of basic emotions. In addition, corpora with a large number of documents have been annotated by a small number of annotators, while sometimes a large number of annotators have been working on a small number of documents. It is important to note that none of them are considered a standard in the field and that their applicability depends on the final application.

The corpus presented in this paper has been designed to go beyond the characteristics of existing corpora. The following requirements have been considered in the design of our corpus:

- The corpus should contain in-context sentences instead of isolated sentences.
- The emotional tags assigned to the sentences in the corpus should be based on subjective human evaluations.
- The set of descriptors used in the annotation should be extensive and flexible.
- The corpus should be annotated by a representative number of human annotators.
- The extension of the corpus should be representative.

In the following sections we describe how these requirements were addressed in the design of our corpus.

3 The design of the corpus

EmoTales has been designed to expand our research in the domain of narrative applications on the automatic detection of emotions in texts (Francisco and Gervás

Table 1 Emotion-related text corpora found in the scientific literature: Bestgen (1993), Pang et al. (2002), Customer Review DataSet (Hu and Liu 2004), MPQA Corpus (Wiebe et al. 2005), OPINE (Popescu and Etzioni 2005), Alm (Alm and Sproat 2005), Blogs06 (McDonald and Ounis 2006), Multiple-aspect Restaurant Reviews (Snyder and Barzilay 2007), Multi-Domain Sentiment DataSet (Blitzer et al. 2007), NTCIR Multilingual Corpus (Seki et al. 2007), Aman (Aman and Szpakowicz 2007) and SEMEVAL 2007 (Strapparava and Mihalcea 2008)

Corpus	Scope		Context	Descriptors		
	Material	Annotators		Dimensions	Categories	Others
Bestgen	4 tales	15	Yes	Evaluation + or −		
Pang	700 reviews	2	Yes	Evaluation + or −		
Customer Review DataSet	5 reviews	31..99	Yes	Evaluation + or −		
MPQA	530 docs. (10,657 sents.)	5	Yes	Evaluation + or −		Subjective, objective
OPINE	200 sents.	2	No	Evaluation + or −		
Alm	22 tales	2	Yes		8 basic	
Blogs06	1,00,649 blogs		Yes	Evaluation + − or both		
Multiple-Aspect Restaurant Reviews	4,488 reviews	3..5	Yes	Evaluation 1..5		
Multi-Domain Sentiment DataSet	2,000 reviews		No	Evaluation + or −		
NTCIR Multilingual Corpus	8,528 sents.	3	Yes	Evaluation + or −		Opinion holders, relevant or opinated sentences
Aman	10,000 sents.	4	No		6 basic	
SEMEVAL 2007	250 headls.	6	No	Evaluation + or −	6 basic	

2006, 2007, 2008). In our previous work we created a small corpus of 8 tales that was annotated by 15 evaluators. The annotation was done in a very rudimentary way. Based on this corpus we created a dictionary of emotions and an approach for the automatic markup of texts with emotions. The new corpus presented here allows us to validate and expand our previous research.

The most important consideration in the design of this new corpus was to have a large emotional corpus annotated by a significant number of annotators. To achieve this goal, the content of the corpus and the subjects were carefully selected and the design of the annotation process was improved.

EmoTales was designed with our final applications in mind, but it was also an attempt to create a corpus that, while as general as possible, facilitated the comparison and evaluation of other resources related to the markup of texts with emotions. We have defined three main steps in this process:

- *Annotation* We propose using both emotional dimensions and emotional categories, including both basic and specific emotions, in the annotation of the corpus. The corpus must also be annotated by a large number of people. For the annotation of EmoTales we had 36 annotators for the corpus annotated with emotional categories and 26 for the corpus annotated with emotional dimensions.
- *Post-processing* The corpus may be post-processed in different ways depending on the intended use. EmoTales can be used without post-processing by considering the initial annotations that were provided by evaluators. Otherwise it can be post-processed to obtain a version of the corpus in which reference values have been identified for each sentence, by selecting the values that were agreed on by most evaluators. These reference values can be found at different specification levels depending on whether a broad set of emotional categories or only basic emotional categories are used.
- *Evaluation* Inter-annotator agreement must be evaluated in order to know how valid the annotations in the corpus are. We need two different metrics to analyze the inter-annotator agreement, one for annotation with emotional dimensions and another for annotation with emotional categories.

3.1 Selection of source texts

Due to our special interest in narrative applications and previous experiences in story generation, we decided to focus our effort on a very specific domain: fairy tales. Fairy tales are generally intended to help children to better understand their feelings, and they usually involve instances of emotions that most people have experienced on their way to maturity (e.g. *happiness, sadness, anger* or *fear*). Emotions in tales, considered from the point of view of a storyteller, have two main functions: to express the personality and internal feelings of a given character at a given moment in the tale, and to induce a certain emotional response in the audience (Kready 1916; Alm et al. 2005). Moreover, tales are especially suitable for the identification and study of emotions because the emotions presented in them are more obvious and explicitly represented than those presented in more complex domains.

Table 2 Distribution of sentences, words and words per sentence (W/S) in the tales chosen

Tale	Author	Sentences	Words	W/S
Cinderella	Brothers Grimm	121	1,079	9
Hansel and Gretel	Brothers Grimm	99	978	10
Rapunzel	Brothers Grimm	104	1,400	13
Sleeping Beauty	Brothers Grimm	66	1,327	20
The Crystal Ball	Brothers Grimm	80	1,084	14
The Emperor's New Suit	H. C. Andersen	151	1,584	10
The Frog Prince	Brothers Grimm	100	1,205	12
The Image of the Lost Soul	Saki	65	891	14
The Lion and the Mouse	Aesop	31	247	8
The Little Match-Seller	H. C. Andersen	70	991	14
The Ox and the Frog	Aesop	25	142	6
The Princess and the Pea	H. C. Andersen	29	373	13
The Selfish Giant	Oscar Wilde	129	1,653	13
The Three Little Pigs	Brothers Grimm	96	1,001	10
The Tortoise and the Hare	Aesop	20	153	7
The Twelve Dancing Princesses	Brothers Grimm	108	1,589	15
The Wicked Prince	H. C. Andersen	75	982	13
The Wolf and the Goat	Aesop	20	137	7

We have selected such a specific domain (fairy tales) due to three main factors:

- The narrative has great cultural importance as a means of communicating, exemplifying, transmitting and teaching complex abstract ideas about values and emotions.
- There is a considerable shortage of work in this domain from the point of view of the representation, identification and annotation of emotions.
- The complexity of the emotional information involved in narrative texts is much higher than in those domains that have so far been the focus of research in Sentiment Analysis domains (blogs, news items, opinion pieces …).

In order to create EmoTales we have selected 18 tales of different length, written in English, making a total of 1,389 sentences and 16,816 words. Tales were chosen according to the practical requirements of our applications, but one of our goals was to cover a broad spectrum of styles by having tales from different authors and time periods.

Table 2 shows the author, number of sentences, words and average of words per sentence of each tale contained in the corpus.

3.2 Annotation granularity and representation of emotions

Sentences are the common unit of linguistic communication as they are used to pack together elements that have more relation to one another than to elements in neighboring sentences. Therefore, it seems reasonable to assign a different emotional

content to each sentence and we have decided to consider the sentence as the emotional unit of our corpus. Each sentence in the tales can have an emotion assigned to it.

The decision to annotate EmoTales with both emotional categories and emotional dimensions ensures compatibility with a large number of emotional representations, and makes the corpus useful for a greater number of applications.

For the representation of **emotional categories** we selected 119 categories (along with the term *neutral*[1]) which aim to cover as many emotional connotations as possible in any text. All those emotional categories were chosen while taking different structures into account; they were constructed using the cluster analysis theory explained in Sect. 2.2. The resulting emotional structure was implemented in an ontology of emotions (OntoEmotions), which will be explained in detail in Sect. 5.2.1. It is relevant to mention the problems inherent to having an extensive set of emotional categories. When the number of emotional categories is too large, the agreement between evaluators usually becomes very low. On the other hand, when the set of emotional categories is very reduced, the emotional descriptions in the text become poor and inaccurate. The aim of our corpus is to have the most accurate corpus as possible so we decided to use an extensive set of emotional categories. The post-processing stage that we included in the annotation process provides a way for improving agreement measures over the final corpus without compromising the quality of the annotation. This allows a customized tailoring of the agreement/coverage ratio to the demands of specific applications.

For the representation of **emotional dimensions** in the corpus we selected the three basic dimensions mentioned in Sect. 2.1: *evaluation, activation* and *power*. In order to help the annotators during the assignment of values for each dimension, we used the SAM standard (Lang 1980). This standard consists of nine values per dimension, which describe progressive changes therein (see Fig. 2). The annotators were asked to select the figure or point between figures that best described the emotion in the sentence they were reading; this point is then mapped into an integer between 1 and 9. The SAM system has been used in other works, with results showing low standard deviation and high inter-evaluator agreement (Grimm and Kroschel 2005).

There are two important issues that we took into account when designing the annotation process:

- Sentences are presented in the context of a tale to make it easier for the annotators to find the appropriate emotion. All sentences are shown sequentially so the annotators decide what the emotional sentences are. When a sentence is not considered emotional it can be annotated as *neutral* by annotators.
- We included texts of different lengths while selecting the tales, both traditional and modern. Regarding the duration of the annotation sessions it has been suggested (Osgood 1967) that subjects can withstand one hour of annotation that may result in 400 annotated sentences. In any case, the patience and endurance

[1] The term *neutral* will be used by annotators in those cases where they could not perceive any of the 119 proposed categories as a clearly identifiable emotion. For example, the sentence "the prince said" is usually annotated as *neutral* by most annotators.

of subjects who are not paid rarely extends beyond 400 annotations, and for those annotators who are not colleagues or friends, the maximum number of annotations is certainly much lower, probably about 50. For that reason it was decided to divide the 18 stories in two sets of 9 stories to avoid overloading annotators with an excessive amount of work. In addition, we used a web interface (described below in Sect. 4) that allows subjects to pause the annotation at any time.

4 Annotation of the corpus

The identification and assignment of emotions to a sentence is a subjective task, so each text from the corpus had to be annotated by several annotators in order to reduce annotator bias. A reference value for each sentence could be obtained based on the emotions the annotators assigned to the sentence in a post-processing stage. If the emotions assigned show too much variability between annotators, it is likely that the sentence will not be assigned a reference value.

The annotation of the corpus was carried out with the evaluation tool TRUE, an on-line platform for multimedia testing evaluation (Planet et al. 2008) developed by La Salle (Universitat Ramon Llull). This tool allows its users to annotate text corpora via a web interface. Evaluators can stop the process at any time, then later resume at the exact point where they had paused.

4.1 Annotation with emotional categories

Figure 1 shows a screenshot of the web interface used to annotate the corpus with emotional categories. On the left there is a fragment of a tale highlighting the sentence to be annotated. On the right, all the available emotions for the annotation of the sentence are listed alphabetically.

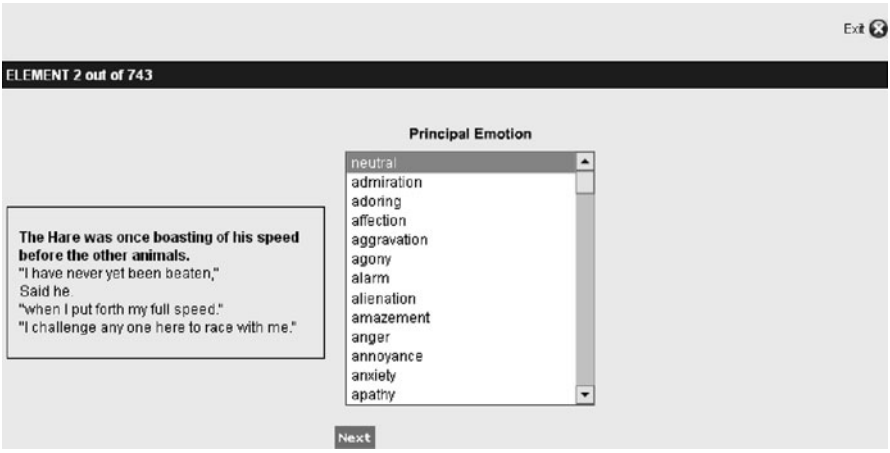

Fig. 1 Screenshot of the web interface for annotating the corpus with emotional categories

Table 3 Number of annotators for each of the tales annotated with emotional categories

Tale	Annotators
The Tortoise and the Hare	12
The Wolf and the Goat	12
The Princess and the Pea	12
The Lion and the Mouse	11
The Little Match-Seller	11
Sleeping Beauty	10
The Twelve Dancing Princesses	10
The Frog Prince	10
The Three Little Pigs	9
Hansel and Gretel	9
The Wicked Prince	9
Cinderella	8
The Selfish Giant	8
Rapunzel	8
The Image of the Lost Soul	8
The Emperor's New Suit	7
The Ox and the Frog	7
The Crystal Ball	7

In order to facilitate the process, annotators were given a list of emotions with the emotional categories grouped semantically to make it faster and easier to find the most suitable emotion for each sentence. The instructions that the annotators were shown as they entered the application for the first time can be seen in Sect. "Appendix 1". Those instructions explained the purpose of the study, and described what emotional categories are.

Annotators were also encouraged to use their first impression and not to try to determine whether there was a single correct answer. An example of the annotation with emotional categories and some recommendations about how to deal with the annotation task were also presented.

Thirty-six annotators participated in the annotation of the corpus with emotional categories, but not all of them annotated all the tales. Each tale was marked up by between 7 and 12 annotators. Table 3 shows the number of annotators per tale. Not all the annotators annotated all the tales because, as mentioned in Sect. 3.2, the patience and endurance of different annotators is not the same.

4.2 Annotation with emotional dimensions

Figure 2 shows a screenshot of the web interface used for annotating the corpus with emotional dimensions. In this screenshot on the left, we see a fragment of a tale where the sentence to be annotated has been highlighted. On the right we can find the SAM standard where annotators must select the point on the scale for each dimension that best represents the emotion transmitted by the sentence.

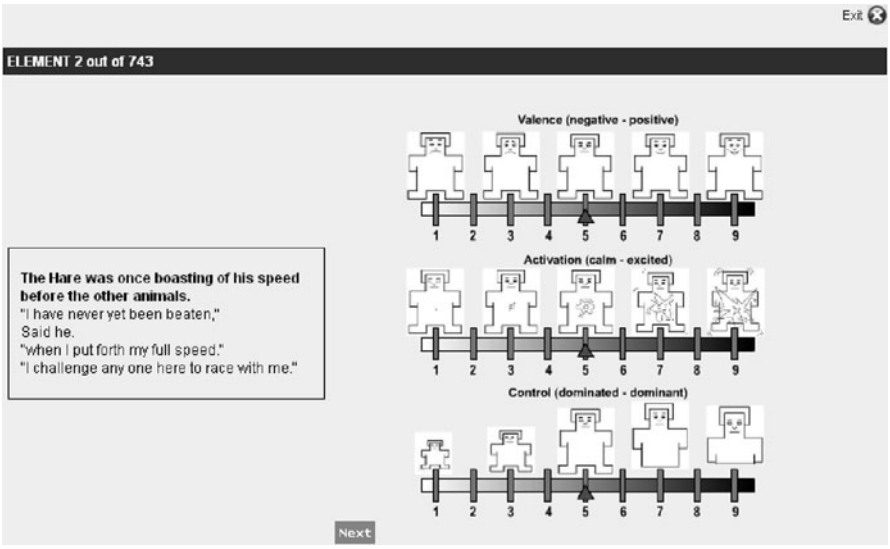

Fig. 2 Screenshot of the web interface for annotating the corpus with emotional dimensions

The instructions that annotators were shown as they entered the application for the first time can be seen in Sect. "Appendix 2". In those instructions they found a presentation of the purpose of the study, and an explanation of what emotional dimensions are, along with a guide about how to use the SAM standard. There were also examples and recommendations to use their first impression as in the emotional categories instructions.

Twenty-six annotators participated in the annotation of the corpus with emotional dimensions, but not all of them annotated all the tales. Each tale was annotated by between 6 and 14 annotators. Table 4 shows the number of annotators per tale.

5 Post-processing of the annotated corpus

In carrying out the annotation of the corpus, we have deliberately given the evaluators a large degree of freedom in deciding what emotional labels to use. The resulting corpus (which was named *View #1*) constitutes a valuable resource inasmuch as it captures the annotating preferences of the various evaluators. As such, it becomes a source for the analysis of which labels might be preferred when annotating in a semi-automatic way.

This freedom in selecting the set of labels to employ introduces an important challenge in terms of agreement between annotators. Where annotators have selected labels of different granularity but related emotional content, consideration of individual annotations in a literal sense will result in disagreement, whereas conceptually, the annotations may well be in accordance. To deal with this problem, we have introduced a stage of post-processing of the corpus at a conceptual level,

Table 4 Number of annotators for each of the tales annotated with emotional dimensions

Tale	Annotators
The Tortoise and the Hare	14
The Princess and the Pea	12
The Selfish Giant	11
The Emperor's New Suit	10
The Wolf and the Goat	7
The Little Match-Seller	7
Sleeping Beauty	6
The Twelve Dancing Princesses	6
The Frog Prince	6
The Three Little Pigs	6
Hansel and Gretel	6
The Wicked Prince	6
Cinderella	6
Rapunzel	6
The Image of the Lost Soul	6
The Ox and the Frog	6
The Crystal Ball	6
The Lion and the Mouse	6

designed to overcome the problem of literal disagreement hiding conceptual agreement. This is done by means of an ontology of emotions.

In the case of emotional categories, conceptual post-processing with the help of an ontology of emotions allows for the generation of more than one view of the annotated resource, each of which relies on labels at a different level of granularity. Two such views are described below: *Subview #2.A* corresponds to the most specific view that can be obtained once literal disagreement has been reduced to conceptual agreement wherever possible, and *Subview #2.B* corresponds to a view of the annotated corpus that uses only labels corresponding to the nine basic emotional categories.

In the case of emotional dimensions, the post-processing stage allows the generation of one more view (*View #3*) which corresponds to that of the annotated corpus in which each sentence has one unique reference value associated for each emotional dimension.

Other types of conceptual post-processing are possible, which allows the conversion of the annotated corpus to views of different granularity in the set of emotional labels contemplated.

In this section we present the different views available in our corpus:

– *View #1*: The set of tales in which each sentence is marked up by several annotators. Each sentence has one emotion assigned per annotator. The emotions assigned to each sentence by different annotators may be different and not necessary the same.

- *View #2*: This view is the result of the post-processing stage in the case of the corpus marked up with emotional categories. As we will explain in Sect. 5.2.2 there are two different ways of performing this post-processing stage, which lead to two different subviews:
  - *Subview #2.A*: As the result of the post-processing stage each sentence in the corpus has a unique emotional category associated, obtained from the different annotations, the most specific emotion with agreementa above a threshold is selected. The emotions associated to each sentence in this view have not been fixed to any level of specification.
  - *Subview #2.B*: As the result of the post-processing stage each sentence has a unique emotional category associated but in this case the level of specification in this category is fixed; only basic emotions appear in this view.
- *View #3*: This view is the result of the post-processing stage in the case of corpus markup with emotional dimensions. As the result of the post-processing stage in this view, each sentence has one unique reference value associated for each of the three emotional dimensions.

5.1 View #1. Corpus with no post-processing

The first view of the corpus is composed of all the sentences in the tales that were marked up by several annotators. In this first view of the corpus no post-processing stage was applied and emotions assigned to each sentence are the original emotions selected by annotators who may have coincided in their selection or not. This first view of the corpus is available for emotional dimensions and emotional categories.

However, there might be some disagreement among annotators, which could be undesirable depending on the expected use of the corpus. In most cases it is useful to have a reference value for each sentence based on annotators' choices. In the next section we explain how this value was obtained in a post-processing stage, first for the corpus annotated with emotional categories, and then for the one annotated with emotional dimensions. As the result of the post-processing stage we obtain a new view of the corpus in which each sentence in the corpus has a single emotion associated, obtained from the annotations made by the evaluators.

5.2 View #2. Reference value for the corpus annotated with emotional categories

To obtain the reference value for each sentence in the case of emotional categories we had different options. The first one was to determine the emotion chosen by most annotators, considering as valuable for the corpus only the sentences where there was agreement among more than half of annotators.

The second option was to minimize the number of sentences with lack of agreement by identifying cases where the specific nominal categories chosen by the annotators, though different, corresponded to related categories that might be grouped together under a more generic category. This procedure required some

means for representing the set of categories involved as a hierarchy of interrelated categories. This seemed a worthy problem to address considering that the percentage of sentences in which half plus one annotators did not agree with the chosen emotion was very significant (in some tales up to 40%). For these cases we implemented an ontology of emotions. This ontology allowed us to relate different emotional categories in order to know if they were similar or not.

Using this ontology we took two different approaches in order to find a reference value for each sentence. In the first case, the level of specification chosen for the reference value only depended on the initial annotations, and the level of specification of emotions assigned to sentences from the same tale could be different. In the second approach, all the emotions selected by the annotators were considered to compute more generic basic emotions, and one of those basic emotions, the most specific one with agreement above a threshold, was selected as a reference value for each sentence. Therefore, in the case of emotional categories we have two subviews for the corpus: one in which the emotions associated with each sentence may belong to different levels of specification (Sect. 5.2.2) and another in which the emotions associated with each sentence are basic (Sect. 5.2.3)

5.2.1 Ontology of emotions

We developed an ontology of emotional categories, called OntoEmotion, as a useful resource for the management of emotional content. By using this ontology we can identify relations between different levels of specification for the same emotion when the emotional content is represented as emotional categories. We took emotional categories (i.e. emotion-denoting words such as *happiness, sadness* and *fear*) as "first class citizens" of our ontology.

5.2.1.1 Basic emotions Based on cluster analysis theory we structured the emotions into clusters. The intention was to integrate the cluster approaches explained in Sect. 2.2. The first step in structuring the emotions was to decide what the basic emotions in our ontology (the different clusters in our approach) would be. As was concluded in Ortony and Turner (1990), researchers cannot identify the basic emotions and we did not even have a satisfactory criterion for identifying basic emotions that is generally acceptable to emotion theorists. However, we tried to find a set of basic emotions in order to create an ontology which allows for the comparison of certain emotions with others. To achieve this we asked ourselves the questions suggested by Ortony and Turner (1990): "What exactly do we mean with basic emotions? In what sense are we using the word 'basic'? What would we do with them if we had them?". The answers to those questions were the following:

- For us basic emotions are superordinate emotions such as *sadness* which subsume other more specific emotions such as *grief* or *despair*.
- The word "basic" is used in the sense of super-ordinate emotion, that is, an emotion that it is not subsumed by any other emotion.
- Once we had our set of basic emotions, our goal was to create a hierarchy of emotions whose roots were the basic emotions. This hierarchy would allow us to

Table 5 Basic emotions selected for cluster analysis approaches presented in Sect. 2.2

Storm and Storm	Shaver et al.	Parrot	Arnold
Sadness	***Sadness***	***Sadness***	***Sadness***
Anger	***Anger***	***Anger***	***Anger***
Fear	***Fear***	***Fear***	***Fear***
Happiness	***Happiness***	*Joy*	
Love	*Affection*	***Love***	***Love***
Disgust			*Aversion*
Anxiety			*Courage*
Contentment			Dejection
Hostility			Desire
Liking			Despair
Pride			Hate
Shame			Hope

make comparisons between different emotions in order to determine whether two emotions are the same, similar or totally different. Two emotions are equal if they are different tags for naming the same abstract emotion. They are similar if they belong to the same cluster (i.e. to the same branch of basic emotions), and they are totally different if they belong to opposite abstract emotions or to two different clusters (i.e. two branches of different basic emotions).

Table 5 shows the basic emotions selected from each of the cluster analysis approaches presented in Sect. 2.2. All these approaches include *sadness, anger* and *fear* as basic emotions so these three emotions were included in our list of basic emotions. When we compare the rest of basic emotions we can see that there are basic emotions that are also shared by all systems. As explained in Ortony and Turner (1990), sometimes the differences between collections of basic emotions are due only to the choice of the tag to refer to the emotion. This is true for *love* and *happiness* emotions, as can be seen in the table; Shaver et al. (1987) refers to *love* as *affection* and Parrott (2001) refers to *happiness* as *joy*. We can conclude that *affection* and *happiness* are common emotions to all the cluster approaches shown in Table 5 and add them to our set of basic emotions with the tags *affection* and *happiness*. *Disgust* is only included in the set of basic emotions of Storm and Storm (1987), but if we look at classical theories of basic emotions we see that *disgust* is included in the set of basic emotions of most of them (Ekman et al. 1982; Izard 1971; Plutchik 1980; Tomkins 1984). Moreover, if we consider basic emotions to be those superordinate ones which include more specific emotions in a hierarchy, we find that *disgust* cannot be included in any of the emotions that are in our current list of basic emotions. It is for these reasons that we include *disgust* as basic emotion in our ontology. *Surprise* is not included in any of the basic emotion sets shown in Table 5, but it is included in most of the classic emotional theories of basic emotions (Ekman et al. 1982; Frijda 1986; Izard 1971; Plutchik 1980; Tomkins 1984), and as in the case of *disgust*, there is no emotion in our current list of basic

Table 6 Subsumption hierarchy for emotions in Table 5

Basic emotions	Specific emotions level 1	Specific emotions level 2
Affection	Liking	
	Lust	Desire
Anger	Rage	Hostility, Hate
Fear	Nervousness	Anxiety
Happiness	Contentment	
	Pride	
	Optimism	Hope
Sadness	Despair	
	Shame	
	Neglect	Dejection

emotions that subsumes the emotion *surprise*; therefore it must also be considered a basic emotion in our ontology.

We also added the term *neutral*, which refers to the lack of emotion, to our set of basic emotions. So far we had defined *sadness, anger, fear, happiness, affection, disgust, surprise* and the additional term *neutral* as basic emotions. To check whether this set was broad enough to include the rest of basic emotions that appear in the approaches shown in Table 5 we placed each of the basic emotions from Table 5 in the clusters obtained from our set of basic emotions. Taking into account the tree approach from Parrott (2001) in which there are three levels of emotions (basic, secondary and tertiary), these emotions were subsumed by our set of basic emotions as presented in Table 6. The only basic emotion contained in Table 5 that could not be subsumed by any of our current basic emotions was *courage*. It was included in our set of basic emotions with the tag *bravery*. This emotion is the opposite of *fear* which was also included as a basic emotion.

In conclusion, we used the following set of basic emotions: *sadness, happiness, surprise, fear, anger, affection, bravery, disgust* and the special term *neutral* (which is not really an emotion, but a tag/concept that represents the absence of emotion).

5.2.1.2 Specific emotions Once the ontology was established with its basic emotions and the different clusters obtained from the classification of approaches in Sect. 2.2, the next step was to complete our ontology by adding those specific emotions that are found in existing emotion literature such as *disappointment, grief, intrigue, melancholy …*

5.2.1.3 Structure of the ontology In OntoEmotion there are concepts that represent language-independent emotions corresponding to common experiences in life. The hypothesis is that we all have the same abstract conception of *Happiness*, for instance, while different words can be used to refer to it. There are also instances in OntoEmotion that represent the words provided by specific languages (e.g. English) for referring to emotions. Therefore, a concept can have multiple instances as a language can give us multiple words to refer to the same

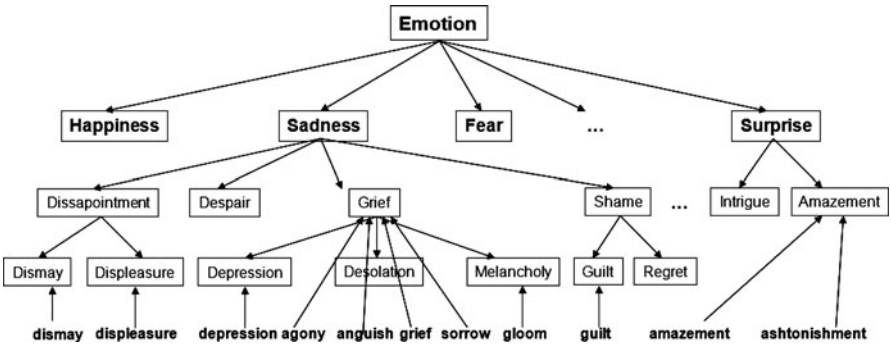

Fig. 3 Fragment of the emotional ontology

emotion. Those instances that correspond to words in a specific language are the ones that were presented to the subjects during annotation.

The root of all emotional concepts in the ontology is the concept *Emotion*. Each emotional concept is a subclass of this root. Emotions are structured in a taxonomy, with the number of levels under each basic emotion depending on the level of available specification for it. For example, *Sadness* has two sublevels of specification. The second level indicates different types of *Sadness*: *Despair, Disappointment, Grief* or *Shame*. Some of these emotions are specialized again in the third level. For example, *Shame* is divided into *Regret* and *Guilt*. On the other hand, *Surprise* only has one sublevel with two emotional concepts: *Amazement* and *Intrigue*

Figure 3 shows a fragment of the ontology. It shows emotional concepts like *Happiness, Sadness, Fear* and *Surprise*. Under those emotional concepts there are instances of these emotional concepts (emotional words) such as *happiness*, *dismay, displeasure* and *depression*.

According to the semantics we chose for our ontology, all the instances of the same emotional concept are synonyms. For example, the words *astonishment* and *amazement* are considered synonyms because both are instances of the emotional concept *Amazement*. Using OntoEmotion we can also obtain the emotion concept directly associated with an emotion word in the ontology, i.e. its parent, as well as with other, more general emotion concepts related to that word, according to the conceptual hierarchy. Finally we can also obtain the synonyms of an emotion word, by noting the siblings of a particular instance.

5.2.2 Subview #2.A. Obtaining a specific reference value from the annotations

To obtain the reference value for each sentence in the case of emotional categories, the two options mentioned before (considering emotions as isolated, unrelated units or considering that emotions selected by annotators might not be the same but related) suggested different treatments.

If more than half of the annotators found no agreement in a sentence, for the first option (view #1) the sentence was initially left with no emotion assigned to it.

Table 7 Example of the assignment of emotions to a sentence by six human annotators [A1..A6]

A1	A2	A3	A4	A5	A6
Agony	Anguish	Grief	Sorrow	Sadness	Sadness

To implement the second option (reference values for each sentence) required a different treatment. Even when less than half of the annotators had agreed on an emotion for a sentence, it is possible they had made decisions that were related and deserved to be taken into account. For these cases we used our emotion ontology in order to find the level of the ontology at which the annotators were agreeing.

In Table 7 there is an example of the assignment of emotions to a sentence by different annotators. In the example, the emotion word *sadness* would be selected as reference value if we take the first option. However, this is not the best choice. The second option involves finding a value in which at least half of the annotators agree by considering the relationships between emotions provided by our ontology. In that case, we find that *agony, anguish, grief* and *sorrow* are synonyms that refer to the same emotional concept *Grief*, so the best assignment for this sentence would be *grief* instead of *sadness*. This second solution, therefore, selects the most specific emotion supported by at least half of the annotators.

The detailed process undertaken to accomplish this task was the following:

1. If at least half of the annotators were in agreement on the assignment of an emotion to the sentence, we took this emotion as reference value for the sentence.
2. Otherwise, we grouped emotions by levels of emotional concepts from the ontology. We obtained all the ancestors of the concepts at the lowest level. If any emotion was supported by at least half of the annotators it was taken as reference value. If there were two emotions that were supported by at least half of the annotators, we took the emotion with the lowest level.
3. We repeated the previous step for each level in ascending order until an emotion supported by most annotators was found.
4. Finally, if there was no emotion that was supported by at least half of the annotators, the sentence in the corpus had no emotion associated.

This process is exemplified in Fig. 4, which shows an example of how to obtain the reference value for a sentence annotated by six evaluators. In the first table we present the assignments initially made for the sentence. The first step is to group the emotions. We obtain the level of each emotion by means of the emotional ontology. The result is shown in the second table. From the second table it can be seen that no emotion is supported by at least half of the annotators, so we obtain the concepts related to the emotions with the lowest level (*Depression, Remorse* and *Helplessness*). We insert all their related concepts in the table (in this case *Grief, Powerlessness, Regret* and *Sadness* three times). The result can be seen in the third table. Based on these results we see that *Sadness* is supported by five annotators, i.e. more than half, so this is the emotion taken as reference value for this sentence.

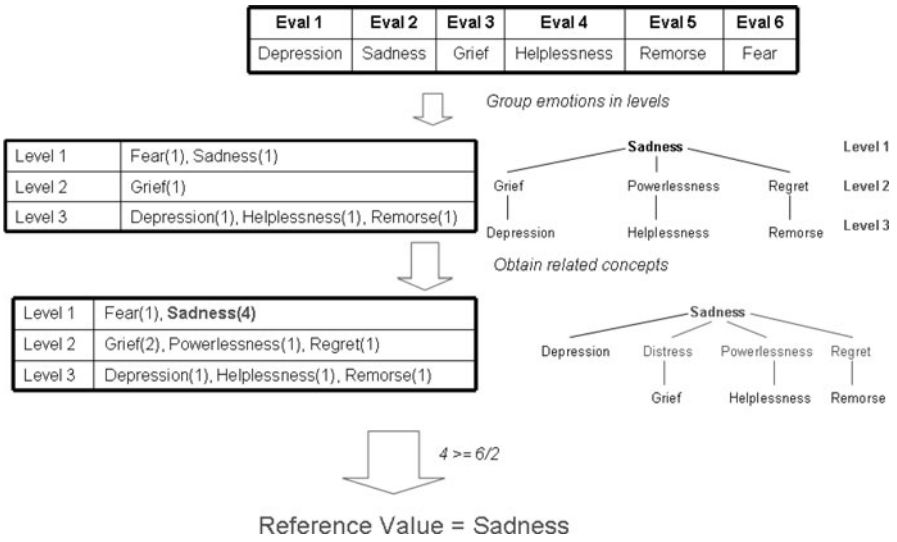

Fig. 4 Example of assignment of the reference value to a sentence in the corpus annotated with emotional categories

Table 8 shows the percentage of sentences from each tale in which final agreement (using the ontology when required) was obtained, and the percentage of these sentences in which it was necessary to use the ontology of emotions to reach agreement. In total there are 1,110 sentences (12,943 words) in which majority agreement was obtained. Table 9 shows the sentence and word count in the corpus for sentences in which majority agreement was obtained and for sentences with lack of agreement among evaluators.

We have empirically evaluated the category reduction performed when the annotators did not agree on the emotion assigned to a sentence. In order to evaluate our reference values we asked annotators whether the re-annotation of the tales that were assigned to them was acceptable. This evaluation was performed by the 13 annotators who had analyzed the most tales in the previous phase. We presented them with pairs <*previous annotation*, *reference value*> along with some of the sentences that they had annotated with the *previous annotation* and which had then been re-annotated by the previously-explained method with the *reference value*. For each pair they had to specify if they agreed with the new annotation or not. Each pair was tested by at least five annotators. The annotators agreed with the new reference value given in 92% of the pairs presented. Therefore it can be considered that the method used to determine the reference value when the annotators did not agree was appropriate. Taking into account this evaluation, it can also be concluded that the ontology used is a valid tool for structuring and relating emotions. The pairs that were identified by the annotators as non equivalent were <*Admiration*, *Happiness*>, <*Arrogance*, *Bravery*>, <*Excitement*, *Happiness*>, <*Gratification*, *Happiness*>, <*Hope*, *Happiness*>, <*Powerlessness*, *Fear*>, <*Suffering*, *Sadness*> and <*Torment*, *Sadness*>.

Table 8 Percentage of sentences with a majority percentage of agreement, and percentage of those sentences in which the ontology was used to obtain agreement

Tale	Agreement %	Ontology use %
Cinderella	85	39
Hansel and Gretel	91	53
Rapunzel	93	33
Sleeping Beauty	53	47
The Crystal Ball	83	53
The Emperor's New Suit	91	18
The Frog Prince	69	3
The Image of the Lost Soul	83	15
The Lion and the Mouse	68	58
The Little Match-Seller	76	57
The Ox and the Frog	96	29
The Princess and the Pea	79	50
The Selfish Giant	80	47
The Three Little Pigs	74	29
The Tortoise and the Hare	60	8
The Twelve Dancing Princesses	81	27
The Wicked Prince	65	40
The Wolf and the Goat	85	33

Once we determined the reference value for each of the sentences, the number of emotional categories, in this view of the corpus, was reduced from 119 to 43. This corresponds to 36% of the total number of categories that were initially available. These 43 emotions are the nine basic emotions (*happiness, sadness, fear, surprise, bravery, affection, anger, disgust* and *neutral*), and the following 34 specific emotions:

admiration	*compassion*	*fury*	*liking*	*powerlessness*	*tenderness*
alarm	*consternation*	*grief*	*loneliness*	*rage*	*unhappiness*
amazement	*decisiveness*	*helplessness*	*longing*	*regret*	*vexation*
anxiety	*enthusiasm*	*hope*	*love*	*relief*	*worry*
arrogance	*excitement*	*humiliation*	*optimism*	*satisfaction*	
care_for	*fright*	*intrigue*	*panic*	*solidarity*	

5.2.3 Subview #2.B. Obtaining a basic reference value from the annotations

As explained in the beginning of Sect. 5, in the case of the corpus markup with emotional categories the second view of the corpus (the view in which the sentences have a reference emotion associated) has two possible subviews: a first subview in which the specification level of the emotions associated to each sentence can be

Table 9 Distribution of sentences, words and words per sentence (W/S) in the final corpus annotated with emotional categories

Tale	Majority agreement			No majority agreement		
	Sentences	Words	W/S	Sentences	Words	W/S
Cinderella	103	887	9	18	192	11
Hansel and Gretel	90	911	10	9	67	7
Rapunzel	97	1,286	13	7	114	16
Sleeping Beauty	35	641	18	31	686	22
The Crystal Ball	66	836	13	14	248	18
The Emperor's New Suit	138	1,425	10	13	159	12
The Frog Prince	69	787	11	31	418	13
The Image of the Lost Soul	54	760	14	11	131	12
The Lion and the Mmouse	21	179	9	10	68	7
The Little Match-Seller	53	727	14	17	264	16
The Ox and the Frog	24	140	6	1	2	2
The Princess and the Pea	23	289	13	6	84	14
The Selfish Giant	103	1,270	12	26	383	15
The Three Little Pigs	71	750	11	25	251	10
The Tortoise and the Hare	12	88	7	8	65	8
The Twelve Dancing Princesses	88	1,261	14	20	328	16
The Wicked Prince	49	611	12	26	371	14
The Wolf and the Goat	17	107	6	3	30	10

different and a second subview in which emotions associated to each sentence are just basic emotions. In this section the second subview is presented.

In order to have not only a corpus marked up with a broad spectrum of emotional categories but also a corpus marked up with basic emotions we obtained another reference value that indicates what the basic emotion is in each sentence. This means the corpus presented in this paper can also be useful for those applications that take only basic emotions into consideration. This new reference value is presented in the second subview of the second view of the corpus.

To obtain this basic reference value the emotions selected by all the annotators are replaced by their related basic emotion from the emotion ontology. After these replacements the basic emotion supported by at least half of the annotators is selected as reference value. If there is no agreement by more than half of annotators in a sentence it is not assigned a reference value.

This process is exemplified in Fig. 5, which shows an example of how to obtain the basic reference value for a sentence annotated by six evaluators. In the first table we present the assignments made to the sentence. The first step is to obtain the basic emotion related to each emotion selected by annotators by means of the emotion ontology. The result is shown in the second table. Looking at the second table it can be seen that *Sadness* is supported by four annotators, i.e. more than half, so this is the emotion taken as basic reference value for this sentence.

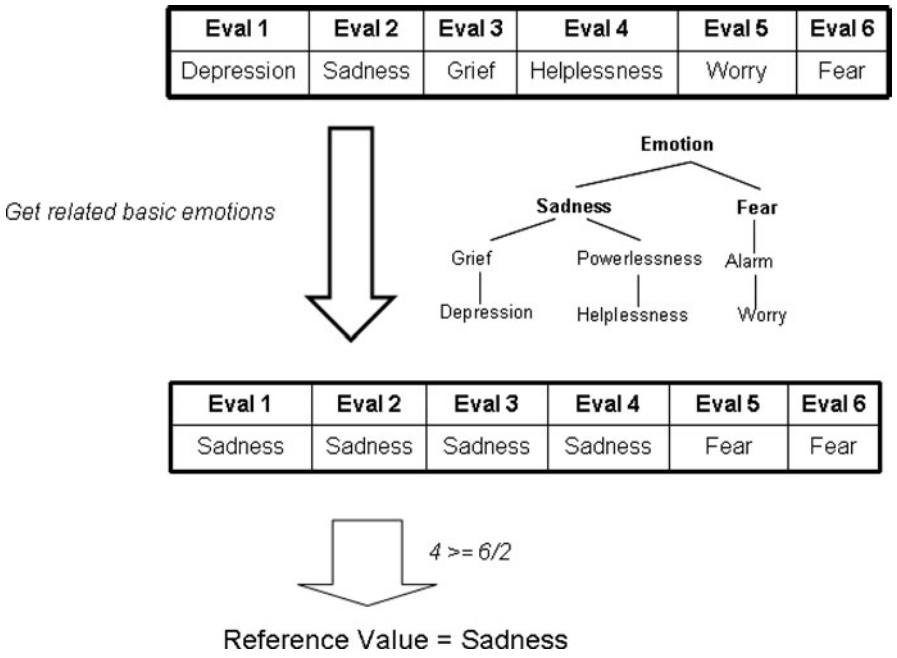

Fig. 5 Example of assignment of the basic reference value to a sentence in the corpus annotated with emotional categories

The percentage of sentences from each tale in which final agreement was obtained is the same as the ones shown in Table 8. If there is majority agreement on a specific emotion this agreement is the same for the basic emotion related to it. To obtain the specific reference value we also used the ontology, therefore if there is no agreement on the specific reference value obtained it means that there are no emotions in the ontology (at no level, including basic emotions) that are supported by at least half of the annotators.

As the result of this post-processing stage we obtained the second subview of the second view of the corpus, composed of 1,110 sentences marked up with nine basic emotions (*sadness, happiness, surprise, fear, anger, affection, bravery, disgust* and *neutral*), plus 279 sentences with no emotion assigned due to the lack of agreement between annotators when assigning an emotion to those sentences.

5.2.4 Characteristics of sentences with no agreement among annotators in the corpus annotated with emotional categories

Looking at Table 9, it can be seen that 70% of sentences with no agreement are sentences with a length above the average of sentences in the same tale. As future work we will undertake an analysis of these sentences to see if they should be subdivided into several sentences or annotated with more than one emotion.

55% of the sentences with no agreement were annotated with emotions belonging to the pairs that were identified by the annotators as non equivalent in the category reduction process (*admiration, arrogance, excitement, gratification, hope, powerlessness, suffering* and *torment*) in Sect. 5.2.2.

In 52% of the sentences without agreement the number of annotators who had selected the majority emotion was equal to the number of annotators who had selected the neutral emotion. It seems that the problem in more than half of the cases was motivated by the lack of agreement when considering if the sentence was emotional or not.

5.3 View #3. Reference value for the corpus annotated with emotional dimensions

In the case of the corpus annotated with emotional dimensions, we obtained the mean score for each emotional dimension in each sentence by obtaining the average value of those assigned to this dimension by the annotators. The result was the value we attached to the sentences in the corpus. Table 10 shows how to get the reference value for a sentence annotated by three evaluators.

To identify the sentences with low agreement we calculated the measurement of polarization (P) (Heise 1970) of the three basic emotional dimensions (evaluation (e), power (p) and activation (a)) for each of the annotators using Eq. 1. P measures the distance between the origin of the three-dimensional space and the particular point under consideration (e, p, a).

$$P = \sqrt{e^2 + p^2 + a^2} \tag{1}$$

We obtained the deviation of this measure with respect to the polarization of the reference value and kept only those sentences in which the mean deviation did not exceed a value of 2.5 (we selected this value because we considered that the sentences with a mean deviation above it were sentences for which agreement is only slight). Table 11 shows the percentage of sentences from each tale assigned a reference value (sentences with agreement between annotators). It can be seen that the percentage of these sentences is slightly higher than in the case of the corpus annotated with emotional categories.

Once we identified those sentences in which the mean deviation was greater than 2.5 we obtained a corpus composed by 1,127 sentences (13,596 words) with values for evaluation, activation and power associated and 262 sentences (3,220 words)

Table 10 Example of reference values for the corpus annotated by human annotators [A1..A3] with emotional dimensions

Dimension	A1	A2	A3	Reference value
Evaluation	6	5	7	6
Activation	8	6	9	7.67
Power	3	2	3	2.67

Table 11 Percentage of sentences taken as valid for our corpus annotated with emotional dimensions

Tale	Sentence agreement %
Cinderella	91
Hansel and Gretel	94
Rapunzel	60
Sleeping Beauty	91
The Crystal Ball	71
The Emperor's New Suit	81
The Frog Prince	82
The Image of the Lost Soul	94
The Lion and the Mouse	84
The Little Match-Seller	71
The Ox and the Frog	56
The Princess and the Pea	97
The Selfish Giant	51
The Three Little Pigs	97
The Tortoise and the Hare	65
The Twelve Dancing Princesses	97
The Wicked Prince	77
The Wolf and the Goat	100

with no values for each dimension associated due to the lack of agreement among annotators. Table 12 shows the sentence and word count in the final corpus for sentences with reference value associated and for sentences with lack of agreement among annotators.

5.3.1 Characteristics of sentences with no agreement among annotators in the corpus annotated with emotional dimensions

Looking at the data in Table 12 it can be seen that 56% of sentences with no agreement are sentences with a length over the average of sentences from the same tale. As in the case of emotional categories in the future we will undertake an analysis of these sentences to see if these sentences should be subdivided into several sentences or annotated with more than one emotion.

22% of the sentences without agreement in the annotation with emotional categories corresponded to sentences without agreement in the annotation with emotional dimensions. In the future we will analyze the semantic content of these sentences to find out the reason for this lack of agreement.

6 Evaluation of the annotated corpus

We evaluated inter-annotator agreement in order to obtain some measurement of the validity of the annotations in the corpus.

Table 12 Distribution of sentences, words and words per sentence (W/S) in the final corpus annotated with emotional dimensions

Tale	Majority agreement			No majority agreement		
	Sentences	Words	W/S	Sentences	Words	W/S
Cinderella	110	972	9	11	107	10
Hansel and Gretel	93	920	10	6	58	10
Rapunzel	62	875	14	42	525	12
Sleeping Beauty	61	1,194	19	5	133	27
The Crystal Ball	57	739	13	23	345	15
The Emperor's New Suit	123	1,263	10	28	321	11
The Frog Prince	82	1,025	12	18	225	12
The Image of the Lost Soul	61	832	14	4	59	15
The Lion and the Mmouse	26	212	8	5	35	7
The Little Match-Seller	50	749	15	20	242	12
The Ox and the Frog	14	82	6	11	60	5
The Princess and the Pea	28	353	13	1	20	20
The Selfish Giant	66	824	12	63	829	13
The Three Little Pigs	94	965	10	3	36	12
The Tortoise and the Hare	14	95	7	7	58	8
The Twelve Dancing Princesses	106	1,554	15	3	35	11
The Wicked Prince	58	741	13	17	241	14
The Wolf and the Goat	20	137	7	0	0	0

6.1 Inter-evaluator agreement for the corpus annotated with emotional categories

We used Fleiss' Kappa statistic (Fleiss 1981) to analyze inter-evaluator agreement in the case of the corpus marked up with emotional categories.

This agreement was computed for all the tales in the corpus. The Kappa statistic was calculated by clustering the emotional categories as they are structured in the ontology presented in Sect. 5.2.1:

Initial Assignment We calculated the agreement for the exact emotional categories used by the annotators.
Merged Synonyms We re-calculated the agreement after all the synonyms of one emotion were merged into one representative value.
Merged Emotions Level 4 We merged emotional categories in the fourth level of the ontology into their corresponding emotional concept from the third level.
Merged Emotions Level 3 We merged emotional categories in the third level of the ontology into their corresponding emotional concept from the second level.
Merged Emotions Level 2 We re-calculated the Kappa statistics after we merged categories in the second level of the ontology into their corresponding basic category.

Table 13 Inter-evaluator agreement for the initial corpus annotated with emotional categories

Tale	Initial Assign.	Merged Synon.	Merged Em. L4	Merged Em. L3	Merged Em. L2
Cinderella	0.21	0.24	0.24	0.25	0.35
Hansel and Gretel	0.14	0.14	0.14	0.15	0.27
Rapunzel	0.19	0.20	0.20	0.20	0.34
Sleeping Beauty	0.12	0.13	0.13	0.14	0.23
The Crystal Ball	0.17	0.17	0.17	0.18	0.28
The Emperor's New Suit	0.25	0.25	0.25	0.25	0.37
The Frog Prince	0.12	0.13	0.13	0.13	0.16
The Image of the Lost Soul	0.20	0.20	0.20	0.20	0.25
The Lion and the Mouse	0.14	0.15	0.15	0.16	0.23
The Little Match-Seller	0.11	0.12	0.12	0.13	0.26
The Ox and the Frog	0.21	0.21	0.21	0.21	0.34
The Princess and the Pea	0.14	0.15	0.15	0.14	0.26
The Selfish Giant	0.17	0.19	0.19	0.20	0.34
The Three Little Pigs	0.12	0.13	0.13	0.13	0.21
The Tortoise and the Hare	0.14	0.14	0.14	0.13	0.34
The Twelve Dancing Princesses	0.16	0.16	0.16	0.18	0.22
The Wicked Prince	0.10	0.10	0.01	0.10	0.15
The Wolf and the Goat	0.13	0.13	0.13	0.13	0.32
Average for the corpus	0.16	0.16	0.16	0.17	0.27

Table 14 Scale for agreement in the annotation of emotional categories

No agreement	Slight	Fair	Moderate	Substantial	Almost perfect
<0	(0..0.20]	(0.21..0.40]	(0.41..0.60]	(0.61..0.80]	=>0.81

Table 13 presents all these results. The average value of the Kappa coefficient for the entire corpus in the case of the categories initially assigned by the annotators is $\kappa = 0.16$ and the average value after the grouping to basic emotions is $\kappa = 0.27$. These levels of agreement are considered *slight agreement* and *fair agreement*, respectively, according to the scale in Table 14 (Landis and Koch 1977).

Table 15 shows the Kappa statistic for the sentences in which the evaluators reached agreement according to the criteria mentioned in Sect. 5.2.2. The average value for the entire database is $\kappa = 0.24$ in the case of the original categories assigned by the annotators and $\kappa = 0.41$ in the case of basic categories. These levels of agreement, which are considered *fair agreement* and *moderate agreement*, according to the scale in Fig. 14, were expected. As was mentioned in Sect. 3.2, having an extensive set of emotional categories implies lower agreement between evaluators. People have different perceptions and interpretations of emotions so it is very difficult to reach substantial agreement. The values obtained are very similar to the agreement levels reported in other work for similar tasks (Devillers et al. 2005;

Busso et al. 2008). It is commonly accepted that the identification of emotions is a subjective task which usually manifests poor inter-evaluator agreement.

6.2 Inter-evaluator agreement for the corpus annotated with emotional dimensions

We have used a comparison between standard deviations of the different annotators to analyze inter-evaluator agreement in the case of the corpus marked up with emotional dimensions.

This agreement was computed for all the tales in the corpus. For each one of the three dimensions we repeated the same process: first, we calculated the average value of the target dimension assigned to each sentence; secondly, we calculated the standard deviation of each annotator in relation to that average value; thirdly, we calculated the average of those standard deviations for each sentence; and finally, we obtained the average of all those numbers assigned to each sentence, which we interpreted as the average deviation of all the annotators when annotating the target dimension in the whole tale.

Table 16 presents the results obtained. The average deviation for the entire corpus in the case of the evaluation dimension is 0.88. In the case of the activation dimension it is higher: 1.07. The power dimension obtains the best result: 0.74. These levels of agreement, according to the scale in Table 17, are considered *substantial agreement* for evaluation and power, and *moderate agreement* for activation. The scale in Table 17 is a hand-made scale which is based on the scale in Table 14, adapting Kappa intervals to average deviation intervals.

Table 18 shows the same analysis for the sentences in which the evaluators reached agreement according to the criteria mentioned in Sect. 5.3, i.e. in which the mean deviation with respect to the polarization of the reference value did not exceed 2.5.

The average values for the final corpus were: 0.68, 0.90 and 0.55 for evaluation, activation and power dimensions, respectively. In this case, the levels of agreement were all considered *substantial*, according to the scale in Table 17.

Compared to the results in Sect. 6.1 these agreement values are clearly higher. This suggests that it is easier for the annotators to agree when using emotional dimensions.

7 Discussion

Regarding the domain of our corpus, fairy tales, it is interesting to point out the role of narrative as a vehicle for exercising and triggering emotions. From this perspective, it makes sense to consider a corpus of narrative texts as a valuable source for exploring what the full range of emotional connotations to be identified from text might be. This choice would also make it possible to link the results of this research to the great effort currently being undertaken by the entertainment industry to explore further uses of information technology in providing new experiences for

Table 15 Inter-evaluator agreement for the sentences with agreement in the corpus annotated with emotional categories

Tale	Initial Assign.	Merged Synon.	Merged Em. L4	Merged Em. L3	Merged Em. L2
Cinderella	0.35	0.37	0.37	0.25	0.41
Hansel and Gretel	0.24	0.25	0.25	0.25	0.42
Rapunzel	0.25	0.27	0.27	0.28	0.39
Sleeping Beauty	0.31	0.32	0.32	0.35	0.40
The Crystal Ball	0.26	0.26	0.26	0.27	0.36
The Emperor's New Suit	0.40	0.37	0.37	0.37	0.49
The Frog Prince	0.15	0.15	0.15	0.16	0.42
The Image of the Lost Soul	0.25	0.25	0.25	0.25	0.42
The Lion and the Mouse	0.24	0.24	0.24	0.32	0.41
The Little Match-Seller	0.21	0.21	0.21	0.24	0.40
The Ox and the Frog	0.24	0.25	0.25	0.27	0.45
The Princess and the Pea	0.27	0.32	0.32	0.32	0.42
The Selfish Giant	0.27	0.31	0.31	0.34	0.52
The Three Little Pigs	0.20	0.22	0.22	0.22	0.43
The Tortoise and the Hare	0.27	0.27	0.27	0.27	0.35
The Twelve Dancing Princesses	0.12	0.12	0.12	0.16	0.36
The Wicked Prince	0.14	0.14	0.14	0.14	0.35
The Wolf and the Goat	0.22	0.22	0.22	0.23	0.40
The whole corpus (in average)	0.24	0.25	0.25	0.26	0.41

gamers and consumers of other interactive media. Where such effort involves the identification, representation, reproduction or induction of emotion in the user, a rich computational representation of emotion, a procedure for attributing emotion to text, and a corpora of material annotated with such a representation would be very valuable resources.

Regarding corpus design, as mentioned in Sect. 2.3, three aspects should be considered in the design of an emotional text corpus: scope, context and descriptors. This section analyses these three aspects in our corpus.

7.1 Scope in EmoTales

The range of emotional classes considered in our corpus has been discussed at length in Sect. 5.2.1.

In our corpus the number of annotators ranged from 6 to 14. The number of annotators used in our corpus may be a sufficient initial step to obtain useful conclusions about the annotation of text with emotions with an average of 9 annotators in the corpus annotated with emotional categories and 7 annotators in the corpus annotated with emotional dimensions. This represents a considerable improvement in the number of annotators with regard to the systems MPQA (Wiebe et al. 2005), OPINE (Popescu and Etzioni 2005), Multiple-Aspect Restaurant

Table 16 Inter-evaluator agreement for the initial corpus annotated with emotional dimensions

Tale	Average deviation		
	Evaluation	Activation	Power
Cinderella	0.85	1.03	0.74
Hansel and Gretel	0.74	0.61	0.69
Rapunzel	1.61	2.18	1.24
Sleeping Beauty	0.77	0.65	0.53
The Crystal Ball	0.76	0.94	0.98
The Emperor's New Suit	0.56	0.39	0.49
The Frog Prince	0.93	0.84	0.62
The Image of the Lost Soul	0.73	0.48	0.40
The Lion and the Mouse	0.79	1.12	0.92
The Little Match-Seller	1.41	1.42	0.84
The Ox and the Frog	0.97	2.89	0.95
The Princess and the Pea	0.91	0.63	0.36
The Selfish Giant	1.60	1.68	0.67
The Three Little Pigs	0.64	0.87	0.78
The Tortoise and the Hare	0.77	0.76	0.93
The Twelve Dancing Princesses	0.45	0.59	0.51
The Wicked Prince	1.07	1.74	0.97
The Wolf and the Goat	0.33	0.49	0.70
The whole corpus (in average)	0.88	1.07	0.74

Table 17 Scale for agreement in the annotation of emotional dimensions based on the average deviation among annotators

No agreement	Slight	Fair	Moderate	Substantial	Almost perfect
>4	(3..4]	(2..3]	(1..2]	(0.50..1]	<=0.51

Reviews (Snyder and Barzilay 2007), NTCIR Multi-lingual Corpus (Seki et al. 2007), SEMEVAL 2007 (Strapparava and Mihalcea 2008) or the systems designed by Pang et al. (2002), Alm and Sproat (2005) and Aman and Szpakowicz (2007) presented in Sect. 2.3. Therefore, the greater number of annotators that participate in the process is an important difference of EmoTales with respect to existing corpora. Bestgen and Alm Bestgen (Bestgen 1993) do not present an ambitious number of annotators and annotated tales, which makes their corpora less valuable when comparing results with those obtained from EmoTales.

All the annotators of EmoTales were fluent English readers, but Spanish native speakers. The choice of non-native speakers as annotators makes this resource dependent on how non-native English speakers perceive emotions in English texts. It has been recognized that language influences the ways in which people think and it has also been noted that emotions depend on national character so we proposed as future work to empirically test if the emotions perceived in English texts by

Table 18 Inter-evaluator agreement for the sentences with agreement in the corpus annotated with emotional dimensions

Tale	Average deviation		
	Evaluation	Activation	Power
Cinderella	0.78	0.88	0.64
Hansel and Gretel	0.67	0.51	0.55
Rapunzel	1.21	2.01	1.14
Sleeping Beauty	0.62	0.56	0.34
The Crystal Ball	0.90	1.06	0.71
The Emperor's New Suit	0.41	0.29	0.29
The Frog Prince	0.85	0.66	0.65
The Image of the Lost Soul	0.51	0.43	0.33
The Lion and the Mouse	0.59	0.85	0.75
The Little Match-Seller	1.13	1.30	0.63
The Ox and the Frog	0.56	2.17	0.28
The Princess and the Pea	0.63	0.58	0.28
The Selfish Giant	1.04	1.16	0.34
The Three Little Pigs	0.55	0.79	0.70
The Tortoise and the Hare	0.10	0.20	0.21
The Twelve Dancing Princesses	0.38	0.53	0.48
The Wicked Prince	1.04	1.80	0.88
The Wolf and the Goat	0.34	0.36	0.65
The whole corpus (in average)	0.68	0.90	0.55

non-native speakers correspond to the emotions that a native English-speaker would notice. The next step would probably be to ensure the reliability of these annotations by repeating the annotation process with native speakers.

7.2 Context in EmoTales

Existing corpora [SEMEVAL 2007 (Strapparava and Mihalcea 2008; Aman and Szpakowicz 2007), or OPINE (Popescu and Etzioni 2005), for example] only contain isolated sentences and the discourse context is not taken into account. One advantage of our corpus is that the sentences are included in a tale in order to contextualize emotions, and the emotional annotations were performed after reading the sentences in the emotional context based on the sequential development of the story.

In the Blogs06 collection (McDonald and Ounis 2006) the discourse context is taken into account as in EmoTales but the plentiful amount of information included is analyzed in terms of date and time of posts, and the occurrences of offensive terms. Therefore, this corpus represents a good source for lexical information, but it is not annotated. The purpose of this resource is to identify spam posts and that is the reason an additional process is required in order to formalize its emotional content.

7.3 Descriptors in EmoTales

Previous attempts have been carried out in recent years to get a text corpus marked up with emotions. Some have been oriented towards the identification of the positive or negative polarity of the texts (Bestgen 1993; Pang et al. 2002; Blitzer et al. 2007; Popescu and Etzioni 2005; McDonald and Ounis 2006) or their classification within a small set of preset emotions (Aman and Szpakowicz 2007; Strapparava and Mihalcea 2008; Alm and Sproat 2005). Our approach goes beyond these attempts by using a broader spectrum of emotional concepts for annotation.

As it was mentioned in Sect. 3.2, when the number of emotional categories is too large the agreement between evaluators usually becomes very low. On the other hand, when the set of emotional categories is very small, the emotional descriptions in the text become poor and inaccurate. The aim of our effort is to have the most accurate corpus as possible so we decided to use an extensive set of emotional categories.

The large set of emotional categories considered in the corpus (which allows us to mark up text with different degrees of specificity) provides a reasonable approach to the emotions observed in the texts. The existing corpora that annotate text with emotional categories tend to use only basic emotions [Alm and Sproat (2005); Aman and Szpakowicz (2007) or SEMEVAL 2007 (Strapparava and Mihalcea 2008)], assuming that annotators will not need a broader range of options to reflect their opinions. However, upon analyzing our corpus we found that annotators used all the 119 categories initially provided. Regarding the use of basic emotions versus the use of specific emotions, annotators preferred the latter. Figure 6 shows the percentage of use of basic, specific and neutral values per tale. The results suggest that the design decision to annotate the text with both basic and specific emotions was a good one. Observing the annotations we see that in 45% of the cases annotators used only specific emotions while in only 15% of the cases they used exclusively basic emotions. Also, the addition of emotional dimensions improves the emotional description of the corpus by capturing supplementary aspects of emotional manifestations such as intensity and variability, and by making it more flexible and adaptable. The corpus is also marked up with emotional dimensions, though not only with the *evaluation* dimension, as in a system such as Bestgen (1993), Pang et al. (2002), MPQA (Wiebe et al. 2005), OPINE (Popescu and Etzioni 2005), Blogs06 (McDonald and Ounis 2006) or SEMEVAL 2007 (Strapparava and Mihalcea 2008), but with the three main emotional dimensions (*evaluation, activation* and *power*).

One of the most interesting aspects of the approach presented in this paper is the post-processing stage. The post-processing stage allows for the selection of more precise tags with higher inter-annotator agreement. This post-processing is an original and promising development in the usual methodology for annotation of emotions. The aim of the post-processing stage was to obtain a reference value that was agreed upon by most evaluators. The reference value obtained was empirically tested by annotators to demonstrate the validity of the post-processing performed.

It is interesting to compare the EmoTales annotation process with the annotation methodology developed for MPQA Corpus (Wiebe et al. 2005). The main goal of

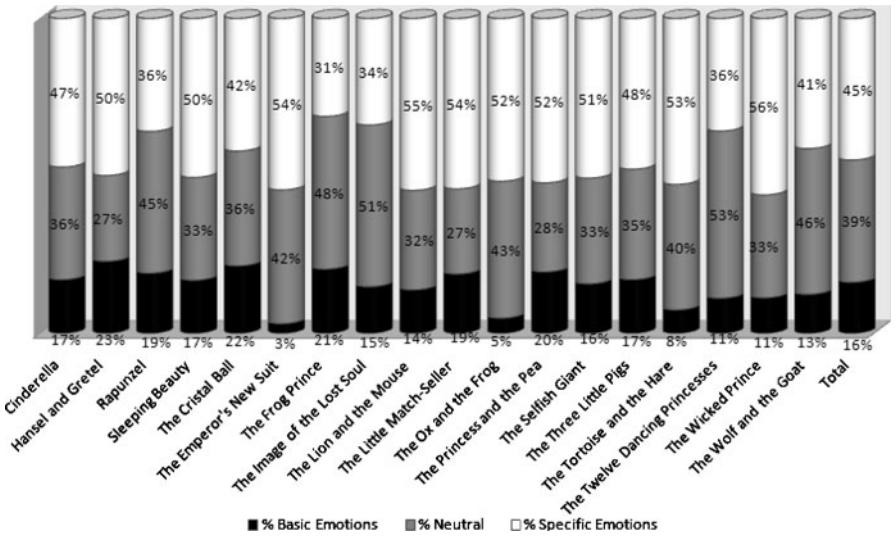

Fig. 6 Percentage of use of basic, specific and neutral emotional values on the annotation of the corpus using emotional categories

the MPQA corpus was to research the use of emotion and opinion in language by using a corpus annotation study. This approach is similar to the one presented in this paper in the sense that the aim is to study the relation between emotions and their representation in texts. However, the MPQA corpus was limited to the annotation of *private states* (Quirk et al. 1985) (internal states that cannot be directly observed). Therefore, not all the texts in the corpus were annotated with emotional information, but only the words and expressions that were considered representative of these private states. In addition, the consideration of a broad set of affective elements as private states (opinions, beliefs, thoughts, feelings, goals, etc.) makes the resource difficult to use when only one of these elements is going to be studied.

8 Conclusions

In this paper a corpus of texts marked up with both emotional categories and emotional dimensions is presented. It contains 1,389 English sentences marked up with emotional categories and emotional dimensions from 18 classic folk tales, and it was annotated by 36 human annotators. First, we presented the design of the corpus and how the annotation process was carried out. Then, a post-processing stage that follows the creation of the initial corpus was discussed. This stage, which is called tag consolidation, aims at finding reference values that were agreed upon by most evaluators. Finally, we evaluated the corpus, obtaining data on inter-annotator agreement and discussing the most relevant aspects of these results.

The initial corpus can be used without post-processing or with a different post-processing treatment better suited to its intended use. The reference values were

obtained based on the annotations made by human subjects. In the case of emotional categories, we carried out the post-processing in two different ways: first, we obtained a specific reference value which indicated an emotion as reference value from all the categories that were presented in the ontology; second, we obtained a basic reference value, that is, a reference value selected from the basic categories. In the case of emotional dimensions we obtained the mean score for each emotional dimension as reference value. As a result we have a very flexible resource that could be useful both to research efforts based on emotional categories (specific or basic) and to efforts based on emotional dimensions.

We evaluated the corpus in order to obtain inter-annotator agreement. This agreement resulted in fair/moderate agreement in the case of emotional categories, and moderate/substantial agreement in the case of emotional dimensions. These results suggest that, although categories are a more flexible and expressive annotation tool, numerical methods reduce the impact of subjectivity in the work of human annotators.

We want to point out that EmoTales illustrates some aspects of the study of how people annotate texts with emotional content. Once the corpus was built we found that it was easy to explore a broad set of research issues that emerged when studying the emotional annotation process of text documents. It is easy to study, for example, what the emotions most and least used by the annotators are, whether there are differences among the annotators' results according to social parameters such as sex, age and education level, what the relation is between emotional dimensions and emotional categories as methods of emotional representation, whether annotators tend to always use the same set of emotional words or not, what the correlation is between the emotional changes in a text document, especially from the narrative point of view, what kind of sentences are considered emotional or what the keywords are that make people identify different emotions in the text, etc. A complete analysis regarding those issues will be addressed in a forthcoming paper.

For evaluation purposes, this corpus may be useful, for instance, in testing different automatic annotation tools using the same set of texts, by using EmoTales as a reference corpus for comparison and evaluation of these software applications. We think the work presented here will be useful to other researchers working on corpus-based approaches to Affective Computing and Natural Language Processing or simply researching models for human communication.

As future work we plan to create similar corpora for other domains and applications, using a higher number of annotators and verifying that all of them are native English speakers. We are also thinking of applying statistical Natural Language Processing or Machine Learning techniques to the corpus in order to obtain information that might help us in the development of a completely automated annotation process. The application of techniques such as keyword spotting and lexical affinity could also help us to obtain an emotional dictionary based on the corpus presented in this paper.

As mentioned in Sect. 5.2.4 another line of future work will be to undertake an analysis of the sentences with no agreement whose length is above the average length of sentences in the same tale to see if they should be subdivided into several sentences or annotated with more than one emotion.

Finally, we consider it interesting to perform some analysis to improve the structure of our ontology in order to find a better position for the annotated emotions that were identified as not equivalent to their reference values by some annotators.

Acknowledgments This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project) and a joint research group grant (CCG08-UCM/TIC-4300) from the Universidad Complutense de Madrid and the Dirección General de Universidades e Investigación of the Comunidad Autónoma de Madrid.We are very grateful to our annotators: Beatriz, Jesús, Lucia, Miguel, Susana, Alaukik, Juan Alvarado, Javier Arroyo, Cristina Arquiaga, Susana Bautista, María del Blanco, Pilar Bravo, Jorge Carrillo, Ana Casas, Alberto Díaz, Borja Foncillas, Ángela Francisco, Patricio Galera, David García, Pilar García, Silvia Garcia, Hector Gómez, Mónica Gónzalez, Francisco Guzman, Nuria Hernández, Jesús Herrera, Guillermo Jiménez, Carlos León, Álvaro Martín, Juanma Martín, Susana Martin, Gonzalo Méndez, Pablo Moreno, Laura Plaza, Celia Pérez, José Ramón Pérez, Patricia Sanz, Cristina Sobrados, Toñi Torreño and Miguel Vázquez. We would also like to thank Pablo Moreno for his useful insights and comments.

Appendices: Instructions given to corpus annotators

Herein we show the instructions presented to the annotators. They were different depending on whether they were annotating the texts using emotional categories or emotional dimensions.

Appendix 1: Annotation using emotional categories

First of all thank you for your collaboration in this study about how people perceive the emotions that are conveyed in tales. Throughout the experiment, the different tales will appear divided into fragments. In each fragment the sentence you must mark at each moment will be highlighted in bold. You must label this sentence using one of the emotions that appear in the list of emotions that are available. In order to identify the beginning of a new story we have put the title of the story in red. Let's try to make clear how this mark up must be done.

The intention is to identify emotions using emotional categories, that is, through the multiple words the language provides us for naming emotions. There are many such words, and a list of emotions is attached as an aid. If you think that a sentence does not convey any emotion you should mark it with the label "neutral".

Please try to do it as fast as possible, and do not take much time thinking about each sentence. The purpose of this experiment is to find out how people interpret the feelings that a tale tries to convey in each sentence; therefore you should check what you interpret. You must use your first impression and not try to determine the "correct" answer or the answer that seems to make more sense because that answer does not exist. For example, the sentence "Cinderella's mother died when she was a child" will be interpreted for some people as sad and the sentence will be marked with the category "sadness". Others will interpret that the sentence conveys grief and they will mark it with the label "grief". And there will be people who feel that this sentence does not attempt to convey anything and will mark it as neutral. None of these three answers is exclusively correct. All are equally valid. What you ought to think when annotating the tale is this: if I were a story-teller and I were reading

this tale, what emotion would I give to each of the sentences to convey the corresponding emotional content to the listener?

Before starting we recommend that you carefully read the list of emotions that you will be provided to get an idea of the possibilities available and where they are located, thus it will be easier and faster to mark up each tale. The list is grouped "semantically". For example, all words that identify sad emotions are in one group. This way it will be easier to find the emotional label you are seeking at any time. You can download this list here.

Keep in mind also that you do not have to mark all the tales at once. You can stop when you want and the point where you have stopped will be registered so you can resume the annotation whenever you want from the point where you stopped last time.

Thank you very much again for your collaboration.

Appendix 2: Annotation using emotional dimensions

First of all thank you for your collaboration in this study about how people perceive the emotions that are conveyed in tales. The aim of this study is to identify emotions by using so-called emotional dimensions. Throughout the experiment the different tales will appear divided into fragments. In each fragment the sentence you must mark at each moment will be highlighted in bold. In order to identify the beginning of a new story we have put the title of the story in red. Let's try to make clear how this mark up must be done.

Each of the three emotional dimensions aims to identify a different type of feeling: joy vs. sadness (evaluation), excitation vs. calm (activation) and control vs. lack of control (power). Your job is to identify the emotion to be conveyed in each sentence in the tale through these three dimensions. If you were a storyteller reading these sentences, would you give the sentence a positive or a negative evaluation? More or less activation? Would you try to transmit control of the situation? In order to help you in this task you will be provided with the SAM scale of figures where you can see how the three emotional dimensions are represented and how this scale is numbered.

The first row of figures represents evaluation, ranging from a smiling figure to a figure with a sad frown. The left end of the scale conveys joy, happiness, satisfaction or optimism. If you feel that the sentence transmits joy it ought to be identified with this part of the scale (9). The other end represents emotions such as sadness, annoyance, dissatisfaction, melancholy, despair or boredom. You can indicate that a sentence attempts to convey sad feelings using the right side of the scale (1). The figures also allow you to describe feelings that are neither entirely happy nor entirely sad by using the figures between the extremes (4, 5, 6, …). If you think that the feeling a sentence is transmitting is between two figures, use the score that appears between them (8, 6, 4, 2) as shown in Fig. 7. This is a scale with a total of nine points and you must choose one of them to identify the evaluation conveyed in each sentence.

The second row of figures represents activation. The right side of the scale indicates that the sentence transmits stimulation, excitement or nervousness. When the sentence suggests nervousness you must select a point on the right (9). If we

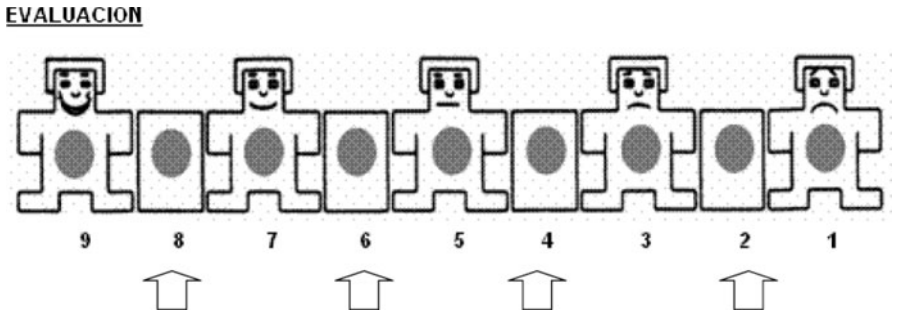

Fig. 7 Intermediate values in the SAM scale

look at the left side of the scale we have the completely opposite feeling, as it conveys relaxation, calm, laziness or sleepiness. To indicate that a sentence conveys calm we will select this extreme of the scale (1). As in the previous scale for evaluation, average levels of arousal or calm can be represented through the figures in the middle of the scale (4, 5, 6, …). Similarly, if you think the excitement or calm transmitted by the sentence is between two figures you must use the numbering between them (8, 6, 4, 2). This is a scale with a total of nine points and you must choose one of them to identify the activation conveyed in each sentence.

The last row represents power. The left side of the scale conveys the feeling of being controlled, guided, intimidated or submissive. To indicate a feeling of total submission this part of the scale (1) will be chosen. The other side conveys control, influence, importance, or self-mastery. Keep in mind that a larger figure represents control and a smaller one represents submission. If you consider that a sentence neither represents control nor total submission you must use the midpoints of the scale (4, 5, 6, …). Again, you can use the scores between two figures if appropriate (8, 6, 4, 2). This is a scale with a total of nine points and you must choose one of them to identify the power conveyed by each sentences.

For example, how can the sentence "I'm angry so leave me alone!" be marked?

- Evaluation: the speaker is angry. Anger involves a negative emotion so we can mark this sentence with the right end of the scale: 1.
- Activation: the anger seems to be a situation that is very active, so we could mark up the sentence with the right side of the scale: 9.
- Power: the speaker is trying to make someone leave him in peace, which leads us to believe that he is trying to dominate someone. So we could mark up the sentence with the right side of the power scale: 9.

To summarize, the annotation for the previous sentence could be the following (although this is not the only possibility; simply one of the options):

- Evaluation = 1
- Activation = 9
- Power = 9

Please try to do it as fast as possible, and do not take too much time thinking about each sentence. The purpose of this experiment is to find out how people interpret the feelings that the tale conveys in each sentence; therefore you should choose your own interpretation of the emotion. You must use the first impression and not try to determine the "correct" answer or the answer that seems to make more sense because that answer does not exist.

Before starting with the first tale, we recommend that you become familiar with SAM; that way it will be easier and faster to mark each of the sentences.

Please also take into account that you do not have to mark all the tales at once. You can stop when you want and the point where you have stopped will be registered so you can resume the annotation whenever you want from the point where you stopped last time.

Again, thank you very much for your collaboration.

References

Alm, C. O., & Sproat, R. (2005). Emotional sequencing and development in fairy tales. In *First international conference on affective computing and intelligent interaction* (pp. 668–674).

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction export find similar. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 579–586).

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In V. Matousek, & P. Mautner (Eds.), *Text, speech and dialogue. Lecture notes in computer science* (Vol. 4629, pp. 196–205). Springer.

Aristotle (1960). *Emotion and personality*. New York: Columbia University Press.

Barreto, A. (2008). *Non-intrusive physiological monitoring for affective sensing of computer users, human computer interaction: New developments*. In K. Asai (Ed.).

Bestgen, Y. (1993). Can emotional valence be determined from words? *Cognition and Emotion, 7*, 21–36.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Ciographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Association for Computational Linguistics*.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*, 335–359.

Campbell, N. (2005). Getting to the heart of the matter: Speech as the expression of affect; rather than just text or language. *Language Resources and Evaluation, 39*, 109–118.

Carbonell, J. (1979). *Subjective understanding: Computer models of belief systems*. PhD thesis, Yale.

Cowie, R., & Cornelius, R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication Special Issue on Speech and Emotion, 40*(1–2), 5–32.

Cowie, R., Douglas-Cowie, E., & Romano, A. (1999). Changing emotional tone in dialogue and its prosodic correlates. In *Proceedings of the ESCA international workshop on dialogue and prosody*, Veldhoven, The Netherlands (pp. 41–46).

Das, S., & Chen, M. (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. In *Proceedings of the Asia Pacific Finance Association annual conference*.

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of WWW* (pp. 519–528).

Devillers, L., Vidrascu, L., & Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks, 18*(4), 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication, 40*(1–2), 33–60.

Evens, M. (2002). New questions for circsim-tutor. In *Symposium on natural language tutoring*, University of Pittsburgh.

Ekman, P., Friesen, W. V., & Ellsworth, P. (1982). What emotion categories or dimensions can observers judge from facial behavior? In P. Ekman (Ed.), *Emotion in the human face* (pp. 39–55). New York: Cambridge University Press.

Fleiss, J. (1981). *Statistical methods for rates and proportions*. New York, USA: Wiley.

Fontaine, J., Scherer, K., Roesch, E., & Ellsworth, P. (2007). The world of emotion is not two-dimensional. *Psychological Science, 18*, 1050–1057.

Francisco, V., & Gervás, P. (2006). Automated mark up of affective information in english texts. In *Text, speech and dialogue* (pp. 375–382). Brno, Czech Republic: Springer.

Francisco, V., & Gervás, P. (2008). Ontology-supported automated mark up of affective information in texts. *Special Issue of Language Forum on Computational Treatment of Language, 34*(1), 23–36.

Francisco, V., & Hervás, R. (2007). Emotag: Automated mark up of affective information in texts. In *EUROLAN 2007 Summer School Doctoral Consortium*, Iasi, Romania (pp. 5–12).

Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.

Grimm, M., & Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. In *IEEE automatic speech recognition and understanding workshop*, San Juan, PuertoRico (pp. 381–385).

Heise, D. R. (1970). The semantic differential and attitude research. In G. F. Summers (Ed.), *Attitude measurement* (pp. 235–253). Chicago: Rand McNally.

Hoffman, T. (2008). Online reputation management is hot but is it ethical? Computer world.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD conference on knowledge discovery and data mining* (pp. 168–177).

Izard, C. E. (1971). *The face of emotion*. New York: Appleton-Century-Crofts.

Kready, L. F. (1916). *A study of fairy tales*. Houghton Mifflin Company. The Riverside Press. http://www.sacred-texts.com/etc/sft/index.htm.

Krenn, B., Pirker, H., Grice, M., Piwek, P., Deemter, K. V., Schröder, M., Klesen, M., & Gstrein, E. (2002). Generation of multimodal dialogue for net environments. In *Proceedings of Konvens*, Saarbrücken, Germany.

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lang, P. (1980). Behavioural treatment and bio-behavioural assessment: Computer applications. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems* (pp. 119–137). Norwood: Ablex Pub. Corp.

Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of intelligent user interfaces (IUI)* (pp. 125–132).

Macdonald, C., & Ounis, I. (2006) The TREC Blogs06 collection: Creating and analyzing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow, 2006.

Merota, G. (2007). Emotional gestures in sport. *Language Resources and Evaluation, 41*, 233–254.

Ortony, A., Clore, G., & Collins, A. (1988). *The cognitive structure of emotions*. New York: Cambridge University Press.

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review, 97*, 315–331.

Osgood, C. (1967). *Contemporary bibliography of research related to the semantic differential technique*. Urbana: University of Illinois Press.

Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Conference on emprirical methods in natural language processing* (pp. 79–86).

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

Parrott, W. (2001). *Emotions in social psychology: Essential readings*. Philadelphia, PA: Psychology Press.

Picard, R. (1997). *Affective computing*. Cambridge: MIT Press.

Picard, R. (2003). Affective computing: Challenges. *International Journal of Human-Computer Studies, 59*, 55–64.

Planet, S., Iriondo, I., Martínez, E., & Montero, J. (2008). TRUE: An online testing platform for multimedia evaluation. In *Second international workshop on EMOTION: Corpora for research on emotion and affect at the 6th conference on language resources and evaluation*.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. In R. Plutchik & H. Kellerman (Eds.), *Emotion: Theory, research and experience: Vol 1. Theories of emotion* (pp. 3–31). New York: Academic Press.

Popescu, A., & Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on human language technology and empirical methods in natural language processing, Association for Computational Linguistics*, Morristown, NJ, USA (pp. 339–346).

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the english language*. Longman: New York.

Russell, J. (1980). A circumflex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Sack, W. (1994). On the computation of point of view. In *Proceedings of AAAI* (p. 1488).

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach (pp. 293–317).

Seki, Y., Kirk Evans, D., Ku, L. W., Chen, H. H., Kando, N., & Lin, C. Y. (2007). Overview of opinion analysis pilot task at NTCIR-6. In *Proceedings of the references 131 workshop meeting of the National Institute of Informatics Test Collection for Information Retrieval Systems* (pp. 265–278).

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology, 52*(6), 1061–1086.

Snyder, B., & Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *Proceedings of the joint human language technology/North American Chapter of the ACL conference* (pp. 300–307).

Storm, C., & Storm, T. (1987). A taxonomic study of the vocabulary of emotions. *Journal of Personality and Social Psychology, 53*.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on applied computing* (pp. 1556–1560). New York, NY, USA: ACM.

Tomkins, S. S. (1984). *Affect theory. Approaches to emotion* (pp. 163–195).

Turney, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics* (pp. 417–424).

Turney, P., & Littman, M. (2003). Measuring praise and criticism: Interference of semantic orientation from association. *ACM Transactions on Information Systems, 21*(4), 315–346.

Wiebe, J. (1994). Tracking point of view in narrative. *Computational Linguistics, 20*, 233–287.

Wiebe, J., & Bruce, R. (1995). Probabilistic classifiers for tracking point of view. In *Proceedings of the AAAI spring symposium on empirical methods in discourse interpretation and generation* (pp. 181–187).

Wiebe, J., Bruce, R., & OHara, T. (1999). Development and use of a gold standard data set for subjectivity classifications. In *Proceedings of the Association for Computational Linguistics* (pp. 246–253).

Wiebe, J., & Rapaport, W. (1988). A computational theory of perspective and reference in narrative. In *Proceedings of the Association for Computational Linguistics* (pp. 131–138).

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

Witt, A., Heid, U., Sasaki, F., & Sérasset, G. (2009). Multilingual language resources and interoperability. *Language Resources and Evaluation, 43*(1), 1–14.

Wright, A. (2009). Our sentiments, exactly. *Communications of the ACM, 52*(4), 14–15.

Yu, H., & Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the conference on empirical methods in natural language processing*.