Emotional Perception of Fairy Tales: Achieving Agreement in Emotion Annotation of Text

Ekaterina P. Volkova[1,2], **Betty J. Mohler**[2], **Detmar Meurers**[1], **Dale Gerdemann**[1], **Heinrich H. Bülthoff**[2]

[1] Universität Tübingen, Seminar für Sprachwissenschaft
19 Wilchelmstr., Tübingen, 72074, Germany
[2] Max Planck Institute for Biological Cybernetics
38 Spemannstr., Tübingen, 72076, Germany

Abstract

Emotion analysis (EA) is a rapidly developing area in computational linguistics. An EA system can be extremely useful in fields such as information retrieval and emotion-driven computer animation. For most EA systems, the number of emotion classes is very limited and the text units the classes are assigned to are discrete and predefined. The question we address in this paper is whether the set of emotion categories can be enriched and whether the units to which the categories are assigned can be more flexibly defined. We present an experiment showing how an annotation task can be set up so that untrained participants can perform emotion analysis with high agreement even when not restricted to a predetermined annotation unit and using a rich set of emotion categories. As such it sets the stage for the development of more complex EA systems which are closer to the actual human emotional perception of text.

1 Introduction

As a first step towards developing an emotion analysis (EA) system simulating human emotional perception of text, it is important to research the nature of the emotion analysis performed by humans and examine whether they can reliably perform the task. To investigate these issues, we conducted an experiment to find out the strategies people use to annotate selected folk fairy tale texts for emotions. The participants had to choose from a set of fifteen emotion categories, a significantly larger set than typically used in EA, and assign them to an unrestricted range of text.

To explore whether human annotators can reliably perform a task, inter-annotator agreement (IAA) (Artstein and Poesio, 2008) is the relevant measure. This measure can be calculated between every two individual annotations in order to find pairs or even teams of annotators whose strategies seem to be consistent and coherent enough so that they can be used further as the gold-standard annotation suited to train a machine learning approach for automatic EA analysis. A resulting EA system, capable of simulating human emotional perception of text, would be useful for information retrieval and many other fields.

There are two main aspects of the resulting annotations to be researched. First, how consistently can people perceive and locate the emotional aspect of fairy tale texts? Second, how do they express their perception of text by means of annotation strategies? In the next sections, we address these questions and provide details of an experiment we conducted to empirically advance our understanding of the issues.

2 Motivation and Aimed Application

Most existing EA systems are implemented for and used in specific predefined areas. The application field could be anything from extracting appraisal expressions (Whitelaw et al., 2005) to opinion mining of customer feedback (Lee et al., 2008). In our case, the intended application of the EA system predominantly is emotion enhancement of human-computer interaction, especially in virtual or augmented reality. Emotion enhancement of

Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 98–106,
Los Angeles, California, June 2010.

computer animation, especially when it deals with spoken or written text, is primarily done through manual annotation of text, even if a rich database of perceptually guided animations for behavioral scripts compilation is available (Cunningham and Wallraven, 2009). The resulting system of our project is meant to be a bridge between unprocessed input text (generated or provided) and visual and auditory information, coming from the virtual character, like generated speech, facial expressions and body language. In this way a virtual character would be able to simulate emotional perception and production of text in story telling scenarios.

3 Related Work

Although EA is often referred to as a developing field, the amount of work carried out during the last decades is phenomenal. This section is not meant as a full overview of the related research as that scope is too great for the length of this paper. To contextualize the research presented in this paper we focus on the projects that inspired us and fostered the ideas.

The work done by Alm (Alm and Sproat, 2005; Alm et al., 2005; Alm, 2008) is close to our project in its sprit and goals. Alm, (2008) aims at implementing affective text-to-speech system for storytelling scenarios. An EA system, detecting sentences with emotions expressed in written text is a crucial element for achieving this goal. The annotated corpus was composed of three sets of children's stories written by Beatrix Potter, H. C. Andersen, and the Brothers Grimm.

Like Liu et al. (2003), Alm (2008) uses several emotional categories, while most research in automatic EA works with pure polarities. The set of emotion categories used is essentially the list of *basic emotions* (Ekman, 1993), which has a justified preference for negative emotion categories. Ekmann's list of basic emotions was extended by Alm, since the emotion of surprise is validly taken as ambivalent and was thus split into *positive surprise* and *negative surprise*. The EA system described in Alm et al. (2005) is machine learning based, where the EA problem is defined as multi-class classification problem, with sentences as classification units.

Liu et al. (2003) have combined an emotion lexicon and handcrafted rules, which allowed them to create affect models and thus form a representation of the emotional affinity of a sentence. Their annotation scheme is also sentence-based. The EA system was tested on short user-composed text emails describing emotionally colored events.

In the research on recognizing contextual polarity done by Wilson et al. (2009) a rich prior-polarity lexicon and dependency parsing technique were employed to detect and analyze subjectivity on phrasal level, taking into account all the power of context, captured through such features as *negation*, *polarity modification* and *polarity shifters*. The work presents auspicious results of high accuracy scores for classification between neutrality and polarized private states and between negative and positive subjective phrases. A detailed account of several ML algorithms performance tests is discussed in thought-provoking manner. This work encouraged us to build a lexicon of subjective clues and use sentence structure information for future feature extraction and ML architecture training.

Another thought-provoking work by Polanyj (2006) shows the influence of the context on subjective clues. This is relevant to our project since we are collecting lexicons of subjective clues and the mechanisms of contextual influence may prove to be of value for future automatic EA system training.

Bethard et at. (2004) provide valuable information about corpus annotation for EA means and give accounts on the performance of various existing ML algorithms. They provide excellent analysis of automatic extraction of opinion proposition and their holders. For feature extraction, the authors employ such well-known resources as WordNet (Miller et al., 1990), PropBank (Kingsbury et al., 2002) and FrameNet (Baker et al., 1998). Several types of classification tasks involve evaluation on the level of documents. For example, detecting subjective sentences, expressions, and other opinionated items in documents representing certain press categories (Wiebe et al., 2004) and measuring strength of subjective clauses (Wilson et al., 2004). All these and many more helped us to decide upon our own strategies, provided many examples of corpus collection and annotation, feature extraction and ML techniques usage in ways specific for the EA task.

4 Experimental Setup

Having established the research context, we now turn to the questions we investigate in this paper: the use of an enriched category set and the flexible annotation units, and their influence on annotation quality. We describe the experiment we conducted and its main results. Each participant performed several tasks for each session. The first task always was a cognitive task on emotion categories taken outside the fairy tales context. The results are discussed in Sections 4.1 and 4.2. The next assignment discussed in Section 4.3 was to annotate a list of words for their inherent polarities. The third task was to read the text out loud to the experimenter. This allowed the participant to feel immersed into the story telling scenario and also get used to the text of the story they were about to annotate for the full set of emotion categories. The annotation process is described in Section 4.4. The last exercise was to read the full fairy tale text out loud again, with the difference that this time their voice and face were recorded by means of a microphone and a camera. The potential importance of the extra data sources like speech melody and facial expressions are further discussed in Section 8 as future work.

Ten German native speakers voluntarily participated in the experiment. The participants were divided into two groups and each participant worked on five of the eight texts. The fairy tale sets for each group overlapped in two texts, which allowed us to achieve a high number of individual annotations in a short amount of time and compare the performance of people working on different sets of texts (see Table 1). Each participant annotated their texts in five sessions, dealing with only one text per session. The fatigue effect was avoided as no annotator had more than one session a day.

User	Fairy Tale ID							
	JG	D	R	BR	FH	DS	BM	SJ
A_1	●	●	●	●	●			
A_2	●	●	●	●	●			
A_3	●	●	●	●	●			
A_4	●	●	●	●	●			
A_5	●	●	●	●	●			
A_6				●	●	●	●	●
A_7				●	●	●	●	●
A_8				●	●	●	●	●
A_9				●	●	●	●	●
A_{10}				●	●	●	●	●

Table 1: Annotation Sets

Positive	Negative
Entspannung (relief)	Unruhe (disturbance)
Freude (joy)	Trauer (sadness)
Hoffnung (hope)	Verzweiflung (despair)
Interesse (interest)	Ekel (disgust)
Mitgefühl (compassion)	Hass (hatred)
Überraschung (surprise)	Angst (fear)
Zustimmung (approval)	Ärger (anger)

Table 2: Emotion Categories Used in the Experiment

4.1 Determining Emotion Categories

First, we needed to define the set of emotions to be used in the experiment. Based on the current emotion theories from comparative literature and cognitive psychology (Ekman, 1993; Auracher, 2007; Fontaine et al., 2007), we compiled a set of fifteen emotion categories: seven positive, seven negative, and neutral (see Table 2). We chose an equal number of negative and positive emotions, since in our experiment the main focus is on the freedom and equality of choice of emotion categories. We aimed at the set to be comprehensive and we also expected the participants to be able to detect each of the emotions in the text as well as express them through speech melody and facial expressions.

The polarity of each category was determined experimentally. Participants were asked to decide on the underlying polarity of each emotion category and then to evaluate each emotion on an intensity scale [1:5], '5' marking extreme polarization, '1' being close to neutral. All participants were in full agreement concerning the underlying polarity of the emotions in the set, while the numerical values varied. It is important to note, that the category *Überraschung (surprise)* was stably estimated as *positive*. In English the word *surprise* is reported to be ambivalent (Alm and Sproat, 2005), but we found that in German its most common translation is clearly positive.

4.2 Emotion Categories Clustering

In the second part of the experiment we asked participants to organize the fifteen emotions into clusters. Each cluster was to represent a situation in which

Cluster	Polarity
{relief, hope, joy}	positive
{joy, surprise}	positive
{joy, approval}	positive
{approval, interest}	positive
{disgust, anger, hatred}	negative
{fear, despair, disturbance}	negative
{fear, disturbance, sadness}	negative
{sadness, compassion}	mixed

Table 3: Emotion Clusters

German Title	English Title	Abbr.
Arme Junge im Grab	Poor Boy in Grave	JG
Bremer Stadtmusikanten	Bremen Musicians	BM
Dornröschen	Little Briar-Rose	BR
Eselein	Donkey	D
Frau Holle	Mother Hulda	FH
Heilige Joseph im Walde	St. Joseph in Forest	SJ
Hund und Sperling	Dog and Sparrow	DS
Rätsel	Riddle	R

Table 4: Stories Used (the titles are shortened)

several emotions were equally likely to co-occur, e.g. a situation formulated by a participant as "*When a friend gives me a nicely wrapped birthday present and I am about to open it.*" was reported to involve such emotions as joy, interest and surprise. On average, each participant has formed 5 clusters with 3–4 items per cluster. The clusters were encoded as sets on unordered pairs of items. Pairs were filtered out if they were indicated by fewer than seven participants. As the result, the following eight clusters were obtained (see Table 3). For most clusters, the categories composing them share one polarity. The {*sadness, compassion*} cluster is the only exception.

It is important to note that the clusters were determined through this cognitive task, independently of the annotations. Since the annotators agree well on clustering the emotions, employing this information captures conceptual agreement between individual annotations even if the specific emotion categories for the same stretch of text do not coincide. However, we intend to keep the full set of emotions for the future corpus expansions.

4.3 Word list Annotation

For each text, we compiled its word list by taking the set of words contained in the text, normalizing each word to its lemma and filtering the set for most common German stop words (function words, pronouns, auxiliaries). Like full story texts, word lists were divided into two annotation sets. At each session, before seeing the full text of the fairy tale, the participant was to annotate each item of the corresponding word list for its inherent polarity. All the words were taken out their contexts and were neutral by default. The annotator's task was to label only those words that had the potential to change the polarity of the context in which they could occur. We purposefully did not limit the task to the words occurring in all texts in order to be able to investigate the stability of participants' decisions. Every annotator worked with five word lists, one for each fairy tale text. The total number of unique items for the first annotation set was 893 words and 823 words long for the second set; 267 and 236 words correspondingly occurred in more than one word list. These words could potentially be marked with different polarity categories, but in fact only about 15% of those words (4% from the total number of items on each of the word lists) were "unstable", namely, labeled with different polarities by the same annotator. The labels received in these cases were either {*positive, neutral*} or {*negative, neutral*}. These words were further "stabilized" by either choosing the most frequent label or the *neutral* label if the unstable word had received only two label instances. The results show that such annotation tasks could be used further for subjective clues lexicon collection.

4.4 Text Annotation

For the third and main part of the experiment, we selected eight Grimm's fairy tales, each 1200 – 1400 words long and written in Standard German (see Table 4). The texts were chosen based on their genre, for in spite of the depth of all the hidden and open references to human psyche and national traditions that were shown in works of (von Franz, 1996; Propp and Dundes, 1977), folk fairy tales are relatively uncomplicated in the plot-line and the characters' personalities. Due to this relative simplicity of the content, we expect the participants' emotional reactions to folk fairy tale texts to be more coherent than to other texts of fiction literature.

The task for the participants was to locate and mark stretches of text where an emotion was to be

conveyed through the speech melody and/or facial expressions if the participant was to read the text out loud. To make the annotation process and its further analysis time-efficient and convenient for both, annotators and experimenters, a simple tool was developed. We created the Manual Emotion Annotation Tool (MEAT) which allows the user to annotate text for emotion by selecting stretches of text and labeling it with one of fifteen emotion categories. The application also has a special mode for word list annotation, where only the three polarity categories are available: positive, negative and neutral. The user can always undo their labels or change them until they are satisfied with the annotation and can submit the results. The main part of the experiment resulted in fifty individual annotations which produced 150 annotation pairs.

5 Analyzing Inter-annotator Agreement

For each of the 150 pairs (two texts annotated by ten annotators, six texts annotated by five annotators), the IAA rate was calculated. However, the calculation of IAA is not as straightforward in this situation as it might seem. In many types of corpus annotation, e.g., in POS tagging, there are previously identified discrete elements. In this experiment we intentionally have no predefined units, even if this makes the IAA calculation more difficult. Consider the following examples:

(1) A_1: "...[the evil wolf]$_X$ ate the girl"
A_2: "...the [evil wolf ate the girl]$_X$"

(2) A_1: "...[the evil wolf]$_X$ ate the girl"
A_2: "...[the evil wolf]$_Y$ ate the girl"

(3) A_1: "...[the evil wolf]$_X$ ate the girl"
A_2: "...the evil wolf ate [the girl]$_X$"

(4) A_1: "...[the evil wolf]$_X$ ate [the girl]$_Z$"
A_2: "...[the evil wolf ate the girl]$_X$"

In example (1) both annotators marked certain stretches of text with the same category *X*, but the annotations do not completely coincide, there is only an overlap. This situation is similar to that in syntactic annotation, where one needs to distinguish between bracketing and labeling of the constituent and measures such as Parseval (Carroll et al., 2002) have been much debated.

Both annotators in example (1) recognize *evil wolf* as marked for *X* and thus this example should be counted towards agreement, while examples (2) and (3) should not. A second type of evaluation arises if the emotion clusters are taken into account. According to this evaluation type, example (2) is counted towards agreement if the categories *X* and *Y* belong to the same cluster.

Example (4) provides an illustration of how IAA is accounted for in a more complex case. Annotator A_1 has marked two stretches of text with two different emotion categories, while annotator A_2 has united both stretches under the same emotion category. Both annotators agree that *the evil wolf* is marked for *X*, but disagree on the emotion category for *the girl*. In order to avoid the crossing brackets problem (Carroll et al., 2002), we treat *the evil wolf ate* as agreement, and *the girl* as disagreement. Although *ate* was left unmarked by one of the annotators, it is counted towards agreement because it is next to a stretch of text on which both annotators agree. Stretches of text the annotators agree or disagree upon also receive weight values: the higher the number of words that belong to open word classes in a stretch, the higher its weight.

The general calculation formulae for the IAA measure are taken from (Artstein and Poesio, 2008):

$$\kappa = \frac{A_o - A_e}{1 - A_e}$$

$$A_o = \frac{1}{i} \sum_{i \in I} arg_i$$

$$A_e = \frac{1}{I^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

A_o is the observed agreement, A_e is the expected agreement, I is the number of annotation items, K is the set of all categories used by both annotators, n_{ck} is the number of items assigned by annotator c to category k.

6 Analyzing Annotation Strategies

Analysis of IAA, presented in Section 5 can answer the first question we aim to investigate: How consistently do people perceive and locate the emotional aspect of fairy tale texts? The second issue necessary for investigation is the annotation strategies people use to express their emotional perception of text. In our experiment conditions, the resulting strategies can be investigated via three aspects: *a*) length of user-defined flexible units *b*) emotional

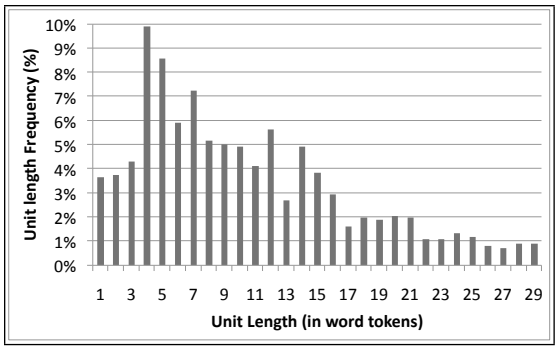

Figure 1: Annotator Defined Unit Length Rating

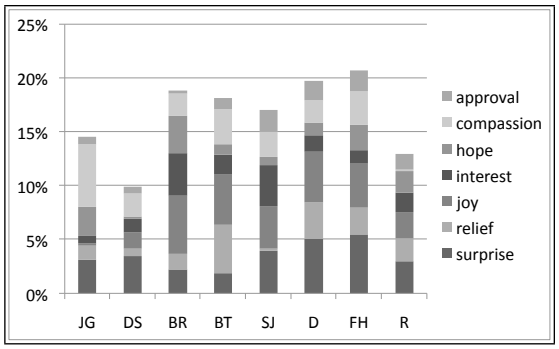

Figure 2: Distribution of Positive Emotion Categories in Texts

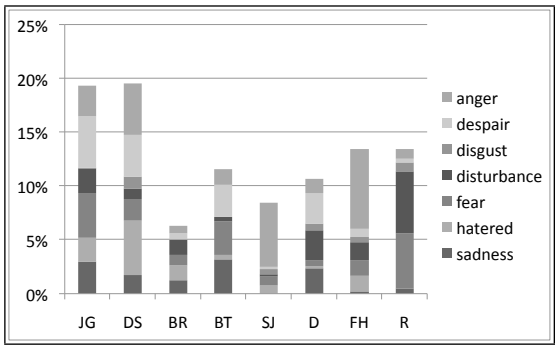

Figure 3: Distribution of Negative Emotion Categories in Texts

composition of fairy tales *c*) emotional flow of the fairy tales. In this section we give a brief account of our findings concerning the given aspects.

The participants were always free to select text stretches of the length they considered to be appropriate for a specific emotional category label. The only guideline they received was to mark the entire stretch of text which, according to their judgement, was marked by the chosen emotion category and, if read without the surrounding context, would still allow one to clearly perceive the applied emotion category label. As Figure 1 shows, the most frequent unit length consists of four to seven word tokens, which corresponds to short phrases, e.g., a verb phrase with a noun phrase argument. We consider the findings to be encouraging, since this observation could be used favorably for the automatic EA system training.

Emotional composition of a fairy tale helps to reveal the overall character of the text and establish if the story is abundant with various emotions or is overloaded with only a few. For our overall research goal, we would prefer the former kind of stories, since they would build a rich training corpus. Figures 2 and 3 give an overview on the average shares various emotion categories hold over the eight texts. It is important to note that 65%– 75% of the text was left neutral. The results show that most stories are rich in positive rather than negative emotions, with two exceptions we would like to elaborate upon. The stories *The Poor Boy in the Grave* and *The Dog and the Sparrow* belonged to different annotation sets and thus no annotator dealt with both stories. These texts were selected partially for their potential overcharge with negative emotions. The hypothesis proved to be true, since the annotators have labeled on average 20% of text with negative emotions, like *hatred* and *sadness*. The only positive emotion category salient for the *The Poor Boy in the Grave* story is *compassion*, which is also mostly triggered by sad events happening to a positive character.

The emotional flow in the fairy tales is illustrated by the graph presented in Figure 4. In order to build it, we used the numerical evaluations obtained in the first part of the experiment and described in section 4.1. For each fairy tale text, each word token was mapped to the absolute value of the average numerical evaluation of its emotional categories assigned by all participants. The word tokens also received its relative position in the text, where the first word was at position 0.0 and the last at 1.0. Thus, the emotional trajectories of all texts were correlated despite the fact that their actual lengths differed. The polynomial fit graph, taken over thus acquired emotional flow common for all fairy tale texts has a wave-shaped form and is similar to the

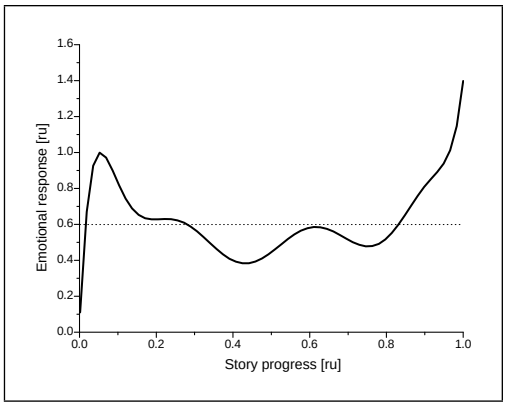

Figure 4: Emotional Trajectory over all Stories

emotional trajectory reported by Alm and Sproat (2005). The emotional charge increases and falls steeply in the beginning of the fairy tale, then cycles though rise and fall phases (which do not exceed in their intensity the average rate of 0.6) and then ascents steeply at the end of the story. We agree with the explanation of such a trajectory, given by Propp and Dundes (1977) and also elaborated by Alm and Sproat (2005) — the first emotional intensity peak in the story line corresponds to the rising action, after the main characters have been introduced and the plot develops through a usually unexpected event. At the end of the story the intensity is highest, regardless whether the denouement is a happy ending or a tragedy. The fact that the fairy tale texts we chose for the experiment are relatively short is probably responsible for the steep peak of intensity in the very beginning of the story — the stories are too short to include a proper exposition. However, we need to investigate further how much of this is a property of texts themselves and how much — the perception (and thus annotation) of emotions.

7 Results

The IAA scores were calculated using the emotion clusters information, for according to the results, participants would often stably use different emotions from same clusters at the same stretch of text.

Four out of ten participants, two from each group (marked gray in Table 1), had very low IAA scores ($\kappa < 0.40$ average per participant), a high proportion of unmarked text, and they used few emotion categories ($<$ 7 categories average per participant), so for the evaluation part their data was discarded. The final IAA evaluation was calculated on all the annotation pairs obtained from the six remaining participants (marked black in table 1), whose average agreement score in the original set of participants was originally higher than 0.50. The total number of annotation pairs amounted to 48: two texts annotated by all the six annotators, six texts annotated by three annotators for each of the two annotation sets.

According to the interpretation of κ by (Landis and Koch, 1977), the annotator agreement was *moderate* on average (0.53), and some pairs approached the *almost perfect* IAA rate (0.83). The IAA rates, calculated on the full set of fifteen emotions, without taking the emotion clusters into consideration, gave a *moderate* IAA rate on average (0.34) and reached *substantial* level (0.62) at maximum. The κ rates are considerably high for the hard task and are comparable with the results presented in (Alm and Sproat, 2005). The word lists have a somewhat lower κ IAA (0.45 on average, 0.72 at maximum), which is due to the low number of categories and the heavy bias towards the *neutral* category. The observed agreement on word lists is considerably high: 0.81 on average, reaching 0.91 at maximum.

While our approach may seem very similar to the one of Alm (2005), there are some important differences. We gave the participants the freedom of using flexible annotation units, which allowed the annotators to define the source of emotion more precisely and mark several emotions in one sentence. In fact, in 39% of all annotated sentences represented a mixture of the *neutral* category and "polarized" categories, 20% of which included more than one "polarized" categories. Another difference is the rich set of emotion categories, with equal number of *positive* and *negative* items. The results show that people can successfully use the large set to express their emotional perception of text (e.g., see Figures 3 and 2).

Other important findings include the fact that short phrases are the naturally preferred annotation unit among our participants and that the emotional trajectory of a general story line corresponds to the one proposed by Propp and Dundes (1977).

8 Future Work

8.1 Corpus Expansion

In the near future, we will expand the collections of annotated text in order to compile a substantially large training corpus. We plan to work further with three annotators that have formed a natural team, since their group has always attained the highest annotation scores for their annotation set, exceeding the highest scores in the other annotation set. The task defined for the three annotators is similar to the experiment described in the paper, with several differences. For the corpus expansion we chose 85 stories by the Grimm Brothers 1400 – 4500 tokens long. We expect that longer texts have more potential space for an emotionally rich plot. Each text will be annotated by two people, the third annotator will tie-break disagreements by choosing the most appropriate of the conflicting categories, similar to the method described by (Alm and Sproat, 2005). It is also probable that a basic annotation unit will be defined and imposed on the annotators, for, as the studies discussed in Section 6 show, short phrases are a language unit most often naturally chosen by annotators.

Each of the annotators will also work with a single word list, compiled from all texts and filtered for the most common stop-words. Each of the words on the word list should be annotated with its inherent polarity (positive, negative or neutral). Since each word on the list is free of its context, the lists provide valuable information about the word and its context interaction in full texts, which can be further used for machine learning architecture training.

We also plan to keep the fifteen emotion categories and their clustering, since it gives the annotator more freedom of expression and simultaneously allows the researches to find the common cognitive ground behind the labels if they vary within one cluster

8.2 Feature Extraction and Machine Learning Architecture Training

When the corpus is large enough, the relevant features will be extracted automatically by means of existing NLP tools, followed by training a machine learning architecture, most probably TiMBL (Daelemans et al., 2004), to map textual units to the emotion categories. It is yet to be determined which features to use, one compulsory parameter is that all the features should be available through automatic processing tools. This is crucial, since the resulting EA system has to be fully automated with no manual work involved.

8.3 Extra Information Sources and their Potential Contribution

We also plan to collect data from other information sources, like video and audio recordings, by inviting amateur actors for story-telling sessions. This will allow emotion retrieval from the speech melody, facial expressions and body language. The manual annotation and the extra data sources can be aligned by means of Text and Speech Aligner (Rapp, 1995), which allows to track correspondences between them. This alignment would most certainly benefit the facial and body animation of the virtual characters, since there is no clear understanding of time correlation between emotions labeled in written text and the ones expressed through speech and facial clues in a story telling scenario. An EA system could also be perfected through a careful analysis of recorded speech and video of story telling sessions — regular recurrence of subjectivity of certain contexts will be even more significant if the transmission of the emotions from the story teller to the listener via mentioned information sources is successful.

9 Conclusions

In this paper, we reported on an experiment investigating the inter-annotator agreement levels which can be achieved by untrained human annotators performing emotion analysis of variable units of text. While EA is a very difficult task, our experiment shows that even untrained annotators can have high agreement rates, even given considerable freedom in expressing their emotional perception of text. To the best of our knowledge, this is the first attempt at emotion analysis that operates on flexible, annotator defined units and uses a relatively rich inventory of emotion categories. We consider the resulting IAA rates to be high enough to accept the annotations as suitable for gold-standard corpus compilation in the frame of this research. As such, we view this work as the first step towards the development of a more complex EA system, which aims to simulate the actual human emotional perception of text.

References

C.O. Alm and R. Sproat. 2005. Emotional sequencing and development in fairy tales. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction (ACII05)*. Springer.

C.O. Alm, D. Roth, and R. Sproat. 2005. Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, volume 2005.

C.O. Alm. 2008. Affect in Text and Speech. *lrc.cornell.edu*.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jan Auracher. 2007. *... wie auf den allmächtigen Schlag einer magischen Rute. Psychophysiologische Messungen zur Textwirkung*. Ars poetica ; 3. Dt. Wiss.-Verl.

C.F. Baker, C.J. Fillmore, and J.B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics Morristown, NJ, USA.

Steven Bethard, Hong Yu, Ashley Thornton, Vasileios Hatzivassiloglou, and Dan Jurafsky. 2004. Automatic extraction of opinion propositions and their holders. In *2004 AAAI Spring Symposium on Exploring Attitude and Affect in Text*, page 2224.

J. Carroll, A. Frank, D. Lin, D. Prescher, and H. Uszkoreit. 2002. Beyond Parseval-Towards improved evaluation measures for parsing systems. In *Workshop at the 3rd International Conference on Language Resources and Evaluation LREC-02., Las Palmas*.

D. W. Cunningham and C. Wallraven. 2009. Dynamic information for the recognition of conversational expressions. *Journal of Vision*, 9(13:7):1–17, 12.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. Timbl: Tilburg memory based learner, version 5.1, reference guide. ilk technical report 04-02. Technical report.

P. Ekman. 1993. Facial Expression and Emotion. *American Psychologist*, 48(4):384–392.

JR Fontaine, KR Scherer, EB Roesch, and PC Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science: a journal of the American Psychological Society/APS*, 18(12):1050.

P. Kingsbury, M. Palmer, and M. Marcus. 2002. Adding semantic annotation to the Penn Treebank. In *Proceedings of the Human Language Technology Conference*, pages 252–256. Citeseer.

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

D. Lee, O.R. Jeong, and S. Lee. 2008. Opinion mining of customer feedback data on the web. In *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, page 230235, New York, New York, USA. ACM.

Hugo Liu, Henry Lieberman, and Ted Selker. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international conference on Intelligent user interfaces - IUI '03*, page 125, New York, New York, USA. ACM Press.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990. Introduction to Wordnet: An online lexical database*. *International Journal of lexicography*, 3(4):235.

L. Polanyi and A. Zaenen. 2006. Contextual valence shifters. *Computing Attitude and Affect in Text: Theory and Applications*, page 110.

V.I.A. Propp and A. Dundes. 1977. *Morphology of the Folktale*. University of Texas Press.

S. Rapp. 1995. Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models. In *Proceedings of ELSNET Goes East and IMACS Workshop*. Citeseer.

M.L. von Franz. 1996. *The interpretation of fairy tales*. Shambhala Publications.

C. Whitelaw, N. Garg, and S. Argamon. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, page 631. ACM.

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

T. Wilson, J. Wiebe, and R. Hwa. 2004. Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the National Conference on Artificial Intelligence*, pages 761–769. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

T. Wilson, J. Wiebe, and P. Hoffmann. 2009. Recognizing Contextual Polarity: an exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399433, September.