Emotional Sequencing and Development in Fairy Tales

Conference Paper *in* Lecture Notes in Computer Science · October 2005

DOI: 10.1007/11573548_86 · Source: DBLP

2 authors:

Cecilia Ovesdotter Alm
Rochester Institute of Technology

Richard Sproat
Google Inc.

Emotional sequencing and development in fairy tales

Cecilia Ovesdotter Alm[1] and Richard Sproat[2]

[1] Dept. of Linguistics, Univ. of Illinois, Urbana, USA, ebbaalm@uiuc.edu
[2] Dept. of Linguistics, Dept. of Electrical Engineering, Univ. of Illinois, Urbana, USA

Abstract. Affect is a transient phenomenon, with emotions tending to blend and interact over time [4]. This paper discusses emotional distributions in child-directed texts. It provides statistical evidence for the relevance of emotional sequencing, and evaluates trends of emotional story development, based on annotation statistics on 22 Grimms' fairy tales which form part of a larger on-going text-annotation project that is also introduced. The study is motivated by the need for exploring features for *text-based emotion prediction* at the sentence-level, for use in expressive text-to-speech synthesis of children's stories.

1 Introduction

This study aims to clarify the relevance of sequential emotion information in text annotation with statistical analysis at the intersentential and text level. In addition, it presents our annotation project and factors that influence the perception of emotions in text.

The study of emotional distributions is motivated by the need for non-lexical features for boosting a classification approach to *text-based emotion prediction*. Predicting emotions in text is useful for subsequent affective text-to-speech synthesis (TTS). We work on child-directed TTS and thus consider fairy tales. Because this genre is schematic, the assumption is that tales have partially predictable *emotional distributions*, e.g. with a NEUTRAL descriptive begin and a HAPPY end. Also, emotions decay, and therefore emotional sentences will most likely be surrounded by NEUTRAL sentences. However, emotions can also be prolonged and even ascend in strength. Intuitively, emotions such as SADNESS appear more likely to span consecutive sentences; while e.g. SURPRISE by nature is more instantaneous. Moreover, the emotional development at the text level can be analyzed from the point of view of the story's emotional trajectory.

In section 2, we present an overview of our on-going text annotation project, and discuss the text-based emotion annotation task based on interannotator agreement measures. Next, we describe the annotated data which was used for statistical analysis in section 3. Then, sections 4 and 5 present statistical analysis of emotional distributions, whereas section 6 concludes with a short discussion and the implications of the study for a learning scenario.

2 Annotation of emotional content in text

Our current text-annotation project targets the annotation of emotional contents in three sets of children's stories, including stories by Beatrix Potter, the Grimms' fairy tales, and tales written by the Danish author H. C. Andersen. The three parts reflect levels of increasing reader maturity. For example, emotions have a notably higher complexity in H. C. Andersen's work, e.g. with embedded emotion assignment where emotions are ascribed to characters by other characters in the story. Eventually, the annotated corpus (approximately 185 stories) will be made publically available.

The annotators are native speakers of US English who attended a literary course on Grimms' Fairy Tales. They mark each sentence in a story with one of 8 primary emotions, listed in table 1, from the point of view of the *feeler* in the sentence, i.e. the emotional target is that perceived by the salient character in the sentence. This helps to focus the annotation task since emotions are often subject to different perspectives.[3] So far, stories have been independently annotated by two people so, as a next step, the annotations produced by each annotator pair are evaluated statistically, and the annotated texts are subsequently post-processed by the first author, who tie-breaks disagreements by choosing the most appropriate of the conflicting labels.

The annotators received a one-hour training session and a manual; however they were not trained together with their eventual pair partner. While such a tandem training scenario may have yielded a higher interannotator score, we did not want to adjust or bias their judgements, but rather get as natural responses as possible, in order to gain more insight into task difficulty and the dynamic nature of emotional targets.

Table 1. Basic emotion categories

Abbreviation	Emotion class
A	ANGRY
D	DISGUSTED
F	FEARFUL
H	HAPPY
Sa	SAD
Su+	POSITIVELY SURPRISED
Su-	NEGATIVELY SURPRISED

Table 2 presents interannotator statistics for a preliminary data set. The two annotators from group A covered a set of 20 Grimms stories (1357 sentences), whereas group B annotated a different set of 22 Grimms stories (1581 sentences). We report on the kappa statistics for inter-annotator agreement, as well as the

[3] Since the narrator may also address the audience directly, the *reader* can also be the *feeler*. Other information is also marked, e.g. emotional intensities, or phrases contributing to emotion assignment.

percent overlap between the paired annotators. For the latter measure, classes were also combined into their semantically intuitive superclasses, where t refers to the top level, i.e. whether a sentence is neutral or non-neutral, m to whether an emotional sentence has neutral, positive or negative valence, and b to the basic level emotion category, e.g. ANGRY, HAPPY, etc., including NEUTRAL.

Table 2. Inter-annotator statistics for 2 groups of paired annotators

	Kappa b	$P(b_{i=j})$	$P(m_{i=j})$	$P(t_{i=j})$
Gr A	0.51	0.64	0.73	0.76
Gr B	0.24	0.45	0.49	0.50

The kappa scores in table 2 are lower than more straight-forward NLP annotation tasks. This confirms the task difficulty and that emotion annotation is sensitive to factors such as subjective interpretation and annotator personality. Also, it is interesting to note that the degree of percent annotation overlap increases as emotion categories are collapsed into more general classes. This could indicate that task difficulty increases for human annotators at the finer-grained distinctions of affect; however, more research is needed to confirm that this is indeed the case.

Moreover, compared to group A, group B showed a substantially lower kappa score on this partial corpus. To investigate the differences in the annotation patterns of group A and B, confusion matrices for both groups were computed which considered all basic emotion categories. The confusion matrices visualize the degree of annotation agreement versus disagreement for an annotator pair. For group A, the diagonal of the left side of table 3 shows that the most frequently

Table 3. Inter-annotation matrix. Group A's a_1 and a_2 and group B's b_1 and b_2 annotated 20 and 22 Grimms' stories, respectively.

a_1/a_2	A	D	F	H	N	Sa	Su+	Su-
A	**87**	0	2	0	4	5	1	8
D	23	2	1	3	7	10	2	8
F	1	0	**56**	2	16	4	2	12
H	3	0	2	**157**	30	1	5	3
N	19	0	16	92	**457**	39	17	41
Sa	6	1	2	2	9	**71**	0	8
Su+	0	2	0	16	14	0	14	4
Su-	8	0	4	5	19	3	5	26

b_1/b_2	A	D	F	H	N	Sa	Su+	Su-
A	*87*	0	6	0	**172**	6	1	4
D	2	1	1	1	**13**	1	0	0
F	3	0	*46*	0	**165**	11	2	4
H	1	0	0	*46*	**192**	1	6	0
N	0	0	4	2	*473*	1	0	0
Sa	1	0	7	3	**137**	*41*	0	7
Su+	0	0	0	3	**65**	1	*10*	1
Su-	4	0	2	0	**41**	1	1	4

chosen category is the same for both annotators for 5 out of 8 emotion labels.[4]

[4] The smallest category, DISGUSTED, appears to be problematic. Within the context of emotion recognition in speech, similarly [6] noted concerns with this category.

For group B, the NEUTRAL category is the dominant source of confusion, as seen in the right side of table 3. This is due to the annotators' sensitivity towards neutral versus non-neutral contents being located at two extremes; b_1 assigned NEUTRAL to 80% of the sentences, whereas b_2 only assigned it to 30%. This highlights that because emotion perception lacks a clear definition as basis and is subjective, NEUTRAL, i.e. the absence of emotion, suffers from the same problem. Nevertheless, the diagonal shows that, besides the NEUTRAL confusion, group B's annotators still had highest within category annotation counts for 6 of the 8 emotion labels.

The distinctions in the kappa results between groups A and B lead us to go one step further and measure intra-annotator consistency. We asked 3 annotators to reannotate 4 Grimms' stories and compared these against the original annotations made by the same annotator, cf. table 4. The results of the intra-

Table 4. Intra-annotator statistics for 3 annotators after reannotating 4 Grimms' stories

	Kappa (b)	$P(b_{i=i})$	$P(m_{i=i})$	$P(t_{i=i})$
Gr A A1	0.64	0.76	0.80	0.82
Gr B A1	0.60	0.67	0.78	0.82
Gr B A2	0.68	0.87	0.88	0.88

annotation experiments in table 4 showed that annotators experienced problems with keeping the emotional targets fixed and stable between annotation times. At this point, it is difficult to say what these fluctuations mean, but they could be related to extra-textual factors, such as deviations in the annotator's mood on the different occasions, cf. [4]. At any rate, these results confirm that emotional targets will tend to be interpreted and perceived dynamically, and that there are bounds on the levels of agreement that can be expected. Further research will help determine what those bounds are.

3 Data for analysis of emotional distributions

The fairy tales used for statistical analysis of emotional distributions consisted of a preliminary tie-broken data set of 22 Grimms' fairy tales or 1580 sentences.[5] The distribution across labels in this subcorpus is included in table 5, with NEUTRAL being most frequent. In this part of the study, we were mostly interested

Table 5. Percent of annotated labels

ANGRY (A): 12%	DISGUSTED (D): 1%	FEARFUL (F): 7%	HAPPY (H): 7%
NEUTRAL (N): 60%	SAD (SA): 7%	POS. SURPRISED (SU+): 3%	NEG. SURPRISED (SU.-): 3%

[5] 1 sentence was an editorial comment and was thus removed from the annotated corpus.

in *local* emotional sequencing. Table 6 shows co-occurrence counts between sentential emotions with emotions of the preceding sentences in the left part, versus with emotions of following sentences on the right. B and E characterized a story's first and last sentence, i.e. when no sentence preceded or followed, respectively.

Table 6. Preceding S emotion counts vs. **following S** emotion counts

0/-1	A	D	H	F	N	Sa	Su-	Su+	B
A	57	4	8	17	90	14	4	1	0
D	2	6	0	0	5	1	0	0	0
H	7	0	13	3	61	7	4	11	1
F	15	0	6	17	50	11	8	4	0
N	94	3	55	55	636	46	12	25	21
Sa	9	1	5	9	58	28	4	2	0
Su-	6	0	1	5	15	4	8	2	0
Su+	3	0	6	3	27	5	1	4	0

0/+1	A	D	H	F	N	Sa	Su-	Su+	E
A	57	2	7	15	94	9	6	3	2
D	4	6	0	0	3	1	0	0	0
H	8	0	13	6	55	5	1	6	13
F	17	0	3	17	55	9	5	3	2
N	90	5	61	50	636	58	15	27	5
Sa	14	1	7	11	46	28	4	5	0
Su-	4	0	4	8	12	4	8	1	0
Su+	1	0	11	4	25	2	2	4	0

4 Statistical analysis of emotional distrubutions

We statistically analyze the following questions:

1. Do fairy tales more frequently begin NEUTRAL and end HAPPY?
2. Does NEUTRAL dominate emotional sentences' immediate context?
3. Which emotions are more often prolonged, and which are not?

The three above points were statistically evaluated by a one-way significance test for difference between the proportions of time that $emotion_i$ and $emotion_j$ immediately preceded or followed $emotion_k$, using the following equation:

$$Z = \frac{\hat{E}_i - \hat{E}_j}{\sqrt{\hat{E}(1-\hat{E})(\frac{2}{n})}} \tag{1}$$

Where $\hat{E}_i = \frac{E_i}{n}$, i.e. the proportion of assignments of $emotion_i$ in sample n, $\hat{E} = \frac{E_i+E_j}{2n}$, and n is the number of sentences with $emotion_k$. Note that this is a particular case of test for difference between 2 proportions when the samples are of the same size.[6] All significance tests were done at 95% confidence level.

The first question considered the prominence of NEUTRAL and HAPPY at opposite story boundaries. Here $emotion_k$ is the special case of *begin* or *end* of a story. The tests showed that NEUTRAL occurred significantly more frequently in the first sentence and HAPPY in the last sentence, compared to other emotions. This confirms that at story boundaries certain emotions can be expected.

[6] Also note that it is only necessary to consider the proportion of an emotion against the proportion of the most frequent one of all other emotions; significance for all smaller proportions follows inductively.

The second question asked, for each emotion, was whether NEUTRAL dominated its adjacent context by testing the difference in proportions between NEUTRAL against other emotions in the preceding and following sentences. For all emotions except DISGUSTED in both contexts and SU- in the following context, NEUTRAL was significantly more frequent compared to other emotions. Thus, for the majority of emotions, NEUTRAL is indeed the most frequent context, which is consistent with the emotional decay factor.

The third question covered emotional prolongation, and was divided into two parts. Given that the above results already demonstrated the dominance of NEUTRAL and the behavior at story boundaries, N, B, and E adjacency counts were excluded.

The first test for emotional prolongation covered whether each emotion was more likely to co-occur with sentences of its own category, as opposed to with other *emotional* categories. Given consecutive sentences, ANGRY and SAD significantly more often preceded and followed themselves compared to other emotions. Moreover, DISGUSTED only significantly more often preceded itself. No statistical evidence was found for HAPPY, FEARFUL and the two SURPRISED conditions. This result corroborates the intuition that some emotions, SURPRISE or FEAR, transition more rapidly than others, such as SADNESS and ANGER. Moreover, a HAPPY emotion expresses contentment with the status quo, whereas negative events force action and keep the narrative plot going, which would explain why HAPPY did not show significant prolongation.

In the second test of emotional prolongation, preceding and following emotions were regarded as either positive PE={H, SU+} or negative NE={A, D, F, SA, SU-}. All NE showed statistical significance for immediate adjacency with NE members. For the positive case, only SU+ was significantly more often followed by PE sentences. Note that it makes intuitive sense that POSITIVE SURPRISE is more notably followed, but not preceded, by positive emotions.

5 Emotional trajectory

Next, we explored if tales reflect a particular emotional trajectory. To increase generalization, emotions were again combined into PE and NE subsets, while N={NEUTRAL} was a singleton class. Each of the 22 stories was divided into fifths and the aggregate frequency counts of PE, NE and N was computed for each fifth over all stories.

Fig. 1 shows that emotional activity in fact corresponded to a wave-shaped pattern. The first fifth is the least emotional level, probably because it descriptively sets the scene of the story. The peak of NE activity in part 2 of the stories could reflect a common pattern in folk tales: due to an event or need early in the story, the hero/heroine must depart on a quest (cf. [5]), and such an event is likely to involve emotional experience. The peak of PE in the final part probably signifies the 'reward' state, i.e. the happy end. Also note that negative emotion increases slightly in the last fifth; according to Propp's schematic outline of folk tales [5], toward the end the villain is 'exposed' as such and then 'punished', i.e.

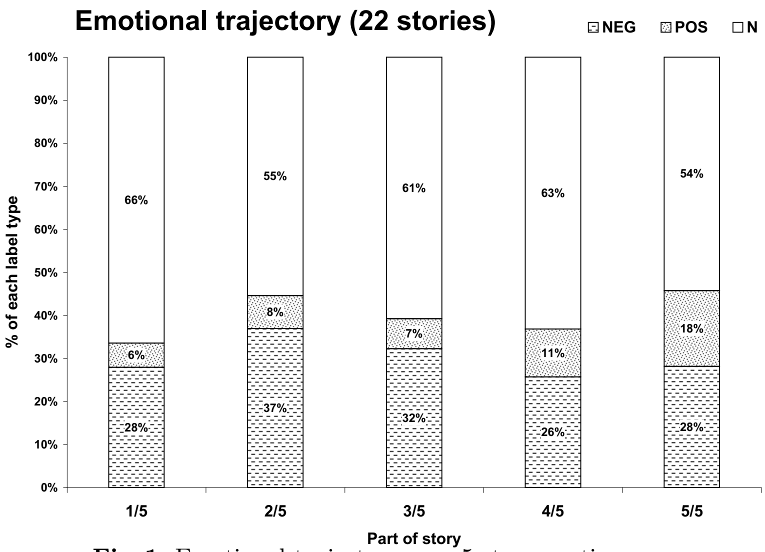

Fig. 1. Emotional trajectory over 5 story portions

events likely to call for negative emotions, such as ANGER. In addition, Propp's template does not always apply, and a few stories end on a negative note.

6 Concluding discussion

This paper discussed emotional distributions in 22 fairy tales in terms of patterns of emotional sequencing and positioning, and also in terms of emotional development, temporally across the story. It additionally provided an overview of the larger on-going annotation project and the issues involved when annotating fairy tale texts for emotional contents.

Although the characteristics and size of the corpus may have influenced the analysis, emotional distributions seem promising as contributing features for *text-based emotion prediction*. In a companion study, we have tentatively observed that annotated sequencing could contribute to slight accuracy improvement of a linear classifier (subject to learner parameters) for bipartite classification of neutral vs. emotional sentences. Naturally, for sequencing to be useful, we must develop automatic inference methods for such features.

Lastly, emotional sequencing could also reflect complex "meta-emotions" [3]. An interesting extension of this work would be to widen the analysis to such cases.

7 Acknowledgments

This work was funded by NSF under award ITR-#0205731, and data annotation by a UIUC Research Board grant. The authors take sole responsibility for the work.

References

1. Cahn, J.: The Generation of Affect in Synthesized Speech. J. Am. Voice I/O Soc. **8** (1990) 1–19

2. van Santen, J. et al.: Applications of computer generated expressive speech for communication disorders. EUROSPEECH (2003) 1657–1660
3. Liu H., Lieberman H., Selker T.: A model of textual affect sensing using real-world knowledge. Intelligent User Interfaces (2003) 125–132
4. Picard, R.: Affective computing. Massachusetts, MIT Press (1997)
5. Propp, V.: Morphology of the Folktale. Austin, Univ. Texas Press (1968)
6. Scherer, K. R.: Vocal communication of emotion: a review of research paradigms. Speech Communication **40:1-2** (2003) 227–256.