AFFECT IN TEXT AND SPEECH

BY

EBBA CECILIA OVESDOTTER ALM

M.A., University of Illinois at Urbana-Champaign, 2002

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Linguistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2008

Urbana, Illinois

Doctoral Committee:

Professor Richard Sproat, Chair
Associate Professor Chilin Shih
Research Assistant Professor Xavier Llorà
Professor Dan Roth
Assistant Professor Roxana Girju

Abstract

As technology and human-computer interaction advances, there is an increased interest in affective computing. One of the current challenges in computational speech and text processing is addressing affective and expressive meaning, an area that has received fairly sparse attention in linguistics. Linguistic investigation in this area is motivated both by the need for scientific study of subjective language phenomena, and by useful applications such as expressive text-to-speech synthesis. The study makes contributions to the study of affect and language, by describing a novel data resource, outlining models and challenges for exploring affect in language, applying computational methods toward this problem with included empirical results, and suggesting paths for further research.

After the introduction, followed by a survey of several areas of related work in Chapter 2, Chapter 3 presents a newly developed sentence-annotated corpus resource divided into three parts for large-scale exploration of affect in texts (specifically tales). Besides covering annotation and data set description, the chapter includes a hierarchical affect model and a qualitative-interpretive examination suggesting characteristics of a subset of the data marked by high agreement in affective label assignments. Chapter 4 is devoted to experimental work on automatic affect prediction in text. Different computational methods are explored based on the labeled data set and affect hierarchy outlined in the previous chapter, with an emphasis on supervised machine learning whose results seem particularly interesting when including true affect history in the feature set. Moreover, besides contrasting classification accuracy of methods in isolation, methods' predictions are combined with weighting approaches into a joint prediction. In addition, classification with the high agreement data is specifically explored, and the impact of access to knowledge about previous affect history is contrasted empirically. Chapter 5 moves on to discuss emotion in speech. It applies interactive evolutionary computation to evolve fundamental parameters of emotional prosody in perceptual experiments with human listeners, indicating both emotion-specific trends and types of variations, and implications at the local word-level. Chapter 6 provides suggestions for continued work in related and novel areas. A concluding chapter summarizes the dissertation and its contributions.

Till min älskade Rubén och vår lilla flicka.

Acknowledgments

Soon after coming back to campus for the Ph.D. program, I realized that a great change was about to take place, which literally has overturned the direction of our department. With Richard Sproat on the Linguistics faculty, students were offered relevant, challenging, motivating, and hands-on courses in computational approaches to linguistics. Moreover, he was sensitive to and interested in listening to the needs of students, and set many of us up with funding and research projects. Having had the privilege to work with Richard for almost 3 years, I can testify to the importance of being blessed with an advisor and mentor who is supportive and encouraging, sincerely cares about students professionally and personally, provides guidance while at the same time respecting one's initiatives and academic creativity, and treats his advisee as a future fellow scholar. I am grateful that Richard introduced me to a general topic and then let me to customize a project which fit my interests. In addition, Richard offered incredible support during the dissertation completion stage, which guided me to complete this manuscript as planned. Interacting with Richard, I also learned first-hand the benefits of efficiency, immediate feedback, and actively seeking out and pursuing research ideas and opportunities.

Moreover, I am greatly in dept to Chilin Shih. Besides giving me untiring feedback, advice, and encouragement, which helped me stay focused and advance, she also always somehow found time to engage into other interesting discussions. These have inspired me to mature as an academic, consider the big picture and my dreams and goals for linguistic research, while maintaining a pragmatic point-of-view. The impact which Chilin has had on my formation and belief system will serve as another source of motivation and guidance as I step beyond my dissertation.

Without Xavier Llorà, I doubt that my vision for exploring speech with evolutionary methods would have become reality. Xavier provided hands-on support and guidance as experimental work was set up and conducted, and I look forward to similar future innovative projects. I am also grateful to David Goldberg, whose class on Genetic Algorithm was an eye-opener for me, and who introduced me to Xavier.

Moreover, I have always admired Dan Roth's willingness to work with students from other fields. An early experience working with his group made me interested in machine learning, spurred me to seek computational

background, and later learn about this area and its implications for and relation to language problems in his courses. I would also like to extend my thanks to Nick Rizzolo, Vasin Punyakanok, and Neelay Shah.

Additionally, thanks to Roxana Girju for stepping into my committee after joining the Linguistics faculty and for providing productive comments and feedback on different work.

This dissertation project was funded by NSF under award ITR-#0205731 and forms part of the *Storyteller* project. I am grateful to colleagues at OHSU and CMU, especially to Jan van Santen and Esther Klabbers.

In addition, I would like to acknowledge educators whose passion helped stir my own, in particular Braj Kachru at UIUC, Christina Bratt Paulston at University of Pittsburgh, Herbert Schendl at University of Vienna, and my history teacher Harald, as well at the support I received from Germanic Languages and Literatures at UIUC, in particular from Shelley, Anna, and Marianne Kalinke.

For having made my time in graduate school more enjoyable, thanks to friends or peers (just to mention a few; Alla, Arthur, Ayelet, Boonpang, Carlos, Evan and Lee, Gaby and Marco, Gabriel, Hahn, Helena, Jennifer, Lisa, Maggie, Nicolas, Oana and Chris, Sarah, Simon, Su-youn, Steve, Tae-jin, and Yuancheng). I would also like to acknowledge those who better deserved to pursue a doctoral degree, but who were taken away too young. Juan Fernando, Tyler, y Victor, les recordaremos.

To my family in law: Raquel y Rubén, Pablo, Carolina y Esteban, y Andrés, muchas gracias por su amor y apoyo.

To my parents Christina and Ove, who are primarily guilty as inspirations for entering academia by living what this is all about during my up-bringing, thanks for your persistent love and care which always brought me forward in life, and for never tiring of listening despite the distance across the pond. I also extend tremendous gratitude to my sister Johanna, her Carl-Johan, and my two nieces Mimi and Lisa. Tack till hela min familj för att ni alltid finns där för oss!

Lastly, I send out my love to Rubén and our (at the point of writing) unborn daughter Mélida Lucia Rubensdotter, who both accompanied me through the dissertation completion stage. Rubén, you always have the greatest belief in me and confidence in that I can achieve anything, and my motivation lies in our life together. Nunca hubiera podido lograrlo sin tu amor.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

"Sorry!" he exclaimed, and his ***voice was troubled with a grief*** *as deep as my own.*

The Shepherd's Story Of The Bond Of Friendship by Hans Christian Andersen (SAD).

"Why this is beautiful, too beautiful to be believed," said the oak ***in a joyful tone.***

The Last Dream Of The Old Oak by Hans Christian Andersen (HAPPY).

Meaning is essential to human language. However, effective communicative messages consist not only of propositional contents, but there are also affective, sociolinguistic, pragmatic, and other meaning components. This dissertation focuses on the affective aspect of meaning and expression in text and speech. The importance of expressive, attitudinal, social/interpersonal, or emotive meaning has been noted by prominent linguistics (Bühler 1934; Lyons 1977; Jakobson 1996; Halliday 1996). However, affect is still an understudied phenomenon in traditional linguistics.

In addition to scientific motivation for having an increased understanding of affect in language and its guises, characteristics, and complexities, there are potentially many useful applications which motivate research in this area. For instance, several applications mentioned by Picard (1997) seem applicable for language data, such as monitoring the affective content of customer service data, or improving human-computer interaction and interface design. Examples of such interfaces are affective chat or email programs, embodied conversational agents, and educational technologies like tutoring systems; also see (Picard 1997). Other areas involve automatically inferring affective orientation of subjective texts, of particular opinions or points of view expressed in text (e.g. for question-answering, summarization, or information extraction, see Wilson, Wiebe, and Hoffman (2005) for a discussion within the context of sentiment analysis), or considering affective meaning and its instantiations within machine translation, literary analysis, language learning, or educational or service-related dialogue systems, and so on. The context of this dissertation project is exploring components which may be useful for automatic storytelling within text-to-speech technology. Such systems may in the future become beneficial for therapeutic education, e.g. for patients with communication disorders that impair the emotional interpretive and productive processing (Picard 1997; van Santen, Black, Cohen, Kain, Klabbers, Mishra, de Villiers, and Niu 2003; Klabbers and van Santen 2004; Loveland 2005).[1] Also, since affect is so important for social interaction between interlocutors, a better understanding of emotional text and speech has broader implications that go beyond creating automatic systems, spanning vastly different fields such as health care, remedial pedagogy, second and foreign language acquisition, marketing, decision-making, politics, and so on.

With state-of-the-art technology, synthesized speech is intelligible to human listeners. However, it generally fails to approach the naturalness of human speech, and a major challenge of natural language systems in general and text-to-speech synthesis in particular is to become more expressive and human. One important component of natural speech is that speakers convey affect by means of verbal as well as paralinguistic cues, such as expressive prosody, facial expressions, gestures, biophysical signals or changes, and posture.[2] Storytelling is complex because in addition to what seems to be a default expressive tone, there may be multiple levels that contribute to storytelling voices, such as affective expressivity, as well as states of more cognitive nature, poetics (e.g. sound symbolism, alliteration, rhyming, repetitions, emphasis, singing, etc.), character impersonation in direct speech (e.g. quoting behavior, gender, age, character or animal type, sociolinguistic speaker factors, etc.), excitement/tension and climax phenomena, and so on.[3] This work's focus is affect, but it also addresses some other aspects in discussions or to minor degree in feature set creation.

A text-to-speech system involves both **text** and **signal** processing (Dutoit 1997; Sproat 1998), and similarly, the linguistic side of affective text-to-speech can be divided into two fundamental problems, which I attempt to address both sides of: automatically predicting affect in text, and characterizing parameters of emotional speech.

On the one hand, generating affective prosody presupposes knowing what affect should be conveyed for a particular part of the input text. In other words, it becomes a text processing task to predict affect labels for and from text. In this dissertation, for predicting affect in text I suggest using multiclass classification.[4] I describe and use a large newly developed corpus of story texts annotated for affect and a hierarchical

[1] Stories are important for children. Alexander, Miller, and Hengst (2001) noted children's social experiences with stories, e.g. children get emotionally engaged and attached to specific stories, build bonds to story characters, and engage into stories' social functions inside and outside the story. Blocher and Picard (2002) reported on tentative positive short-term learning curves and system reception, when children diagnosed with autism or PDD interacted with a system for training emotion recognition via video clips, although some children needed help to engage and interact with it. A neat feature of the ASQ system was the hands-on doll interface which also engaged motor-skills. *Anger* was mostly best recognized, whereas *surprise* and *sadness* were more difficult; perhaps this could also be related to doll design.

[2] Overall, increased attention is being paid to *affective computing* (Picard 1997) within the AI community at large. To illustrate, in 2004 an AAAI symposium was held on the topic of *Exploring Attitude and Affect in Text: Theories and Applications*, in 2005 the *First International Conference on Affective Computing and Intelligent Interaction* took place, and the speech community has organized workshops, special sessions, etc. related to affect in speech.

[3] For example, Reilly and Seibert (2003) discussed evaluative devices from narratives elicited from children. In addition to speech-related cues, they also mentioned gestural and facial behaviors, and narrator's insertion of evaluative comments and quoted speech.

[4] Reducing to binary classification at the top level of the affect hierarchy, see Chapters 3 and 4.

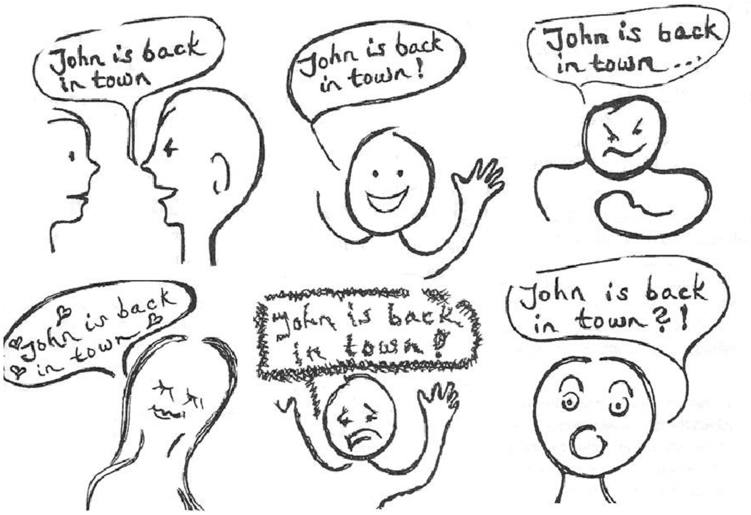


Figure 1.1: Predicting affect in text is an extremely challenging problem because potentially, an utterance or sentence can be rendered affectively in various ways. The question is for what sentences does a particular affect make sense? Illustrations from Alm-Arvius, C. (1998). *Introduction to Semantics.* Lund: Studentlitteratur. (p. 15). Used with permission.

affect model. Predicting affect labels is a highly challenging task, because hypothetically, any utterance or sentence can be rendered with different types of affective inflection. For example, Fig. 1.1 illustrates the utterance *John is back in town* being expressed in different affective tones. The challenging question is for what sentences in a story does a particular affect make sense? This is further complicated by the fact that what is acceptable or appropriate in terms of affective meaning is rather subjective. For instance, perception of affective meaning in text relates to issues such as point-of-view, personal experience and acquired knowledge, contextual factors, and prototypical expectations, which may or may not be shared across individuals. Moreover, another challenge for predicting affect at a particular point in a text is that affect may occur at different scopes or levels in text, such as when considering the more explicit or local emotion, the general or background mood, or the personality of the entire text. However, these scopes or levels co-occur, overlap, and are hard to define. Fig. 1.2 illustrates this with a fuzzy continuum model.[5]

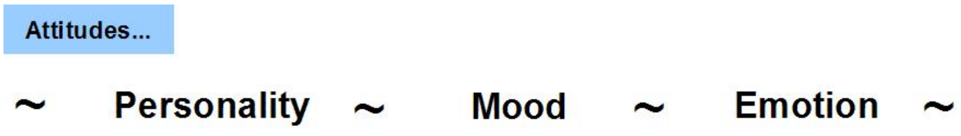


Figure 1.2: A tentative continuum of over-lapping affective scopes in text

On the other hand, expressive signal processing presupposes an understanding of how parameters of the

[5] There may of course also be other scopes of affect in text, for example attitudes.

voice inflection, or *prosody*, convey the appropriate affect. I attempt to approach this problem via interactive evolutionary search for exploring the relationship between affective states and fundamental parameters of speech in resynthesis experiments with human participants using the active interactive Genetic Algorithm (aiGA) by Llorà, Sastry, Goldberg, Gupta, and Lakshmi (2005).

These two problems span different modalities of affect in language; affect in written versus spoken text. Within an expressive text-to-speech system, a text-based affect prediction component can feed a text with automatically inferred and marked up affect labels into a signal processing component that in turn draws on findings from affective speech research to more appropriately render an expressive acoustic signal.

The rest of this work is structured as follows. Ch. 2 surveys various aspects of related previous work. Then, the next two chapters deal with affect in text. Ch. 3 provides a summarizing overview of the newly developed corpus and the annotation project which gathered human-labeled data for the empirical work on automatic affect prediction, covered in Ch. 4. I then move on to empirical work on affect in resynthesized (i.e. computer-modified) speech with evolutionary interactive computation in Ch. 5. Next, I suggest directions for future work in Ch. 6 and, finally, I provide concluding remarks in Ch. 7.

Chapter 2

Literature review

This chapter discusses literature which is related to or inspired work described later in this dissertation. Affect phenomena have fascinated scholar in several fields for a long time. I touch upon several areas, however this is not a complete account of topics and perspectives due to the breath of affect-related work.

The chapter is divided into subsections which describe related work and provide motivation for this study. After a look at how to characterize emotion and various psychological models of emotion, I discuss linguistic work on emotional language, followed by selected pieces on subjective language and affect in computational linguistics, and on affect in speech. Finally, I also give a brief overview of some related insights on stories, oral narratives, and computerized storytelling.

2.1 Characterization of emotion

Affect is a term used for fuzzy notions such as *feelings*,[1] *emotion*, *personality*, *mood* and *attitudes*, and affective states may be real or faked, as well as closely connected to interpersonal interaction, context, rituals or social stratifications, patterns, and expectations (Besnier 1990). This study mostly uses the terms *emotion* and *affect* interchangeably. Similarly slippery, an emotion is transient, intensity-varying, functional, "accompanied by physiological and behavioral changes" (p. 536), neurological at the core, and conveyed by various modalities such as verbal content, prosody, and facial expressions, with channel-preferences influenced by development (Reilly and Seibert 2003). Affect might occur at different levels; for instance, it has been suggested that *emotion* be distinguished from *mood* and other affect levels (Picard 1997; Scherer 2003), e.g. by differences in temporal or intensity resolution or other factors. Moreover, emotion involves both physical and cognitive components (Picard 1997). Ch. 3 more explicitly presents the model of affect used in the present study.

Affect has been given a great deal of attention in psychology, and I will briefly cover my interpretation of a few approaches to emotion (either theories or ways to describe or label emotion): *categorical*, *dimensional*,

[1] These terms may be used differently, e.g. (Picard 1997) separates *feelings* (also relating to e.g. hunger, etc.) from *emotions*.

appraisal, *Jamesian*, *social constructivist*, and *free labeling*.[2] Studies of affect recognition, generation, or labeling by humans or computational agents is generally based on one or more of the presented perspectives and ideas.

In wanting to show that emotional expressions stem from earlier development (and describing them as similar to reflexes occurring also when useless) Darwin (1890) observed "that the same state of mind is expressed throughout the world with remarkable uniformity" (p. 5). He also noted caveats with observing expressions, and that it is hard to describe such states' physical movements. In a similar vein, *categorical* views of emotion may involve assumptions about universality of certain emotions. A compelling observation was that emotive facial expressions in images were cross-culturally recognized well above chance. Ekman and Friesen (1971) investigated *happiness*, *anger*, *sadness*, *disgust*, *surprise* and *fear* in New Guinea, and found the rate at which individuals selected the corresponding emotive faces to match a story/situation significant. An exception was that *fear* was mistaken as *surprise*, interpreted as pointing to the cultural aspect of emotional discrimination.[3] Ekman and Friesen (1971) (p. 70) concluded that

> The results for both adults and children clearly support our hypothesis that particular facial behaviors are universally associated with particular emotions. With but one exception, the faces judged in literate cultures as showing particular emotions were comparably judged by people from a preliterate culture who had minimal opportunity to have learned to recognize uniquely Western facial expressions.[4]

Nevertheless, the Ekmanian view remains controversial. For instance Russell and Fernández-Dols (1997) critiqued the relevance, methods, and rigor of the "Facial Expression Program" for emotion.

Ekman (1994) and others have also discussed the concept of *emotion families*, which share properties, and "constitute a theme and variations" (p. 19). Interpreting this concept, for example *worry*, *anxiety*, and *panic* might be varying intensity levels of *fear*,[5] and the notions of *hot* vs. *cold anger* (Scherer 2003), which have been explored within the speech community, are members of the *angry* emotion family. The notion of emotion family corresponds well with the flexible nature of linguistic categories (Taylor 1995), which

[2] Also see Cornelius (2000) for a stimulating and more in-depth discussion on four theoretical perspectives and their interactions (Darwinian, Jamesian, Cognitive, and Social Constructivist).

[3] The study by Ekman and Friesen (1971) investigated both adults and children; by mistake the *fear* vs. *surprise* distinction was not tested on children. Gendered t-tests or testing for Westernization degree were not significant.

[4] Ekman and Friesen (1971) continued saying that "Further evidence was obtained in another experiment in which the facial behavior of these New Guineans was accurately recognized by members of a literate culture. In that study, visually isolated members of the South Fore posed emotions, and college students in the United States accurately judged the emotion intended from their videotaped facial behavior" (p. 70). They also referred to studies which hinted at blind and seeing children having similar facial behaviors. Moreover, they acknowledged the cultural component in evoking an emotion, influencing consequent actions, and social rules on exhibiting facial behaviors. And Ekman (1994) stated that "[i]ndividual differences in our experience allow for enormous variations in the specifics of what calls forth an emotion that are attributable to personality, family and culture" (p. 16). Work by Ekman and colleagues also lead to the development of the Facial Action Coding System.

[5] The example reflects my perception, which may differ from that of other people.

include both more prototypical as well as more marginal members, as well as fuzzy boundaries influenced by e.g. functional properties (Labov 1973). Scholars disagree on what the set of more canonical emotions (often called *basic emotions*) are, and this view has also been criticized (Goddard 1998).[6] However, in vision and speech research "the Big Six" (Cornelius 2000) including *happiness, fear, anger, surprise, disgust*, and *sadness* appear quite often. Moreover, some believe that basic emotions might blend or sequentially combine into complex emotions. For example, in their work on an affective email client, Liu, Lieberman, and Selker (2003) suggested that *fear* followed by *happiness* represented *relief.* However, some emotional states may not be captured this way. Also, some states are perhaps better described as cognitive (e.g. *love, seriousness, curiosity*), although the line seems challenging to draw.[7]

In a *dimensional* account, emotion is situated in continuous multidimensional space with dimensions such as *evaluation*, *activity*, and *potency* related to semantic differentials (Osgood 1969). Dimensions may also have different descriptive terms, e.g. instead of *activation* the literature may use a similar concept like *arousal* or *intensity*, *pleasantness* or *valence* instead of *evaluation*, *dominance* is related to *power*, and so on. As an example for how dimensions can be used to encode emotion, one might describe *elated happiness* as reflecting extremely high pleasantness and activation, or *rage* extremely high activation but very low pleasantness. There is, however, a lack of consensus on which dimensions to consider, although some reappear frequently. Another issue with dimensions is that gradual rating scales may be subject to diverse interpretations. Circumplex models with emotion words, considering both dimensions and categories, have been suggested by e.g. Plutchik and Russel (Winton 1990). Moreover, the salience of certain emotional dimensions may depend on modality. Discussing confusion patterns in the context of McGurk effect with visual and audio modalities, Fagel (2006) suggested that valence was more informative for the visual mode, whereas arousal applied more to the audio mode.

In the cognitively-oriented field *appraisal* is at the center of emotion. For example, the OCC model (Ortony, Clore, and Collins 1988) links emotions to reasoning and agents' reactions to actions and objects in relation to goals, and whether the action was turned to the agent itself or to someone else. Whereas the OCC model involves context and reasoning, it appears to encompass rule interaction and fairly complex interpretation. For example, knowledge-based approaches to semantics may be brittle or impractical implementation-wise, due to the challenge in enumerating human experience.

The *Jamesian* view sees emotion as experience of physical and bodily responses (Cornelius 2000). An

[6] I make no distinction between basic emotions and superordinate emotion categories, as Schröder (2004) does. Moreover, Ekman has expressed the belief that there are no non-basic emotions, characterizing emotion given certain properties: "automatic appraisal, commonalities in antecedent events, presence in other primates, quick onset, brief duration, unbidden occurrence, and distinctive physiology" as well as "a distinct, universal signal" (p. 18). Not all potential emotions fulfill these.

[7] It could potentially be argued that this also reflects *surprised.*

interesting question is how perception and experience relate, e.g. if interpreting others' emotion is linked to feeling. Keltner and Ekman (2003) touched upon how Oatley suggested that readers experience emotion by cognitive simulation, placing themselves in the position of characters of a narrative and imagining their own emotional reaction. Also, the *social constructivist* point of view regards emotions as shared cultural constructs, depending on factors such as social rules, practices, and values (Goddard 1998; Cornelius 2000). In other words, social analysis and consideration of cultural functions and interpretation seems required to understand emotion. This does not necessarily contradict, but rather complement, a universalist perspective.

Lastly, hefty debate on how to characterize and label affect has led to *free labeling*, criticizing basic emotions for not reflecting adequate affective diversity, and instead advocating the interpretive imagination of individual perceivers. However, applying this approach may result in an impractically large set of different emotion labels and it is challenging to determine a principled way for reclassification into a smaller confusion set. Shrinking a free label set to accommodate this dilemma might even result in sets similar to basic emotions or data set reduction. For example, after grouping items from a large set of open-ended responses to a perception test on characterizing certain fairy tale sentences, Bralé, Maffiolo, Kanellos, and Moudenc (2005) interestingly noted that, although other cases occurred, "Big Six" emotions were frequent in answers. Devillers, Abrilian, and Martin (2005) subsequently shrunk the free labels assigned to emotional TV segments to a label set, and in a study on medical emergency call center agent-client dialogues Vidrascu and Devillers (2005) emphasized non-basic blends and a rich multilevel annotation scheme, also considering dimensions, but subsequently removed blended or conflictingly labeled data, based on assumptions of data-reliability, when classifying emotional speech data.

Overall, this section showed that there are many differences in opinions on how to characterize affect, as well as lack of a clear consensus definition (Picard 1997; Scherer 2003). As a consequence, *neutrality*, i.e. absence of emotion, suffers from a similar dilemma. At the same time, different approaches and views of emotion complement each other and jointly create a basis for understanding the complexities of affective language phenomena. Calling a particular approach "better" would be simplistic.

This study mostly applies a categorical scheme of basic emotions for labeling, classification, generation or evolution. However, the discussions in this dissertation span more than one particular paradigm. For example, the feature set and affect hierarchy used for affect prediction in text to some degree also drew on insights from the dimensional camp, whereas the concern voiced about the importance of knowledge, context, association, and experience in Ch. 3 related to social-constructivist and appraisal models.

2.2 Affect in text-based linguistics

Expressive language is a powerful tool. For example, Hoey (2000) provided an insightful discussion on how Noam Chomsky's (later) writing used persuasive evaluative rhetoric to strengthen his agenda and discourage challenge of his work. Also, since affect extends beyond language, so does research on language and affect. For instance, the field of psychiatry has analyzed language and speech by patients with affective or other disorders, e.g. (Ogilvie, Stone, and Shnediman 1969; Andreasen and Pfohl 1976; Ragin and Oltmanns 1983; Harvey 1983; Fraser, King, Thomas, and Kendell 1986; Hoffman, Stopek, and Andreasen 1986; Ragin and Oltmanns 1987; Calev, Nigal, and Chazan 1989).

As noted in Ch. 1, linguists have long acknowledged expressive meaning and language forms and functions. Nevertheless, affective language phenomena have received comparatively little attention in mainstream linguistics, e.g. (Besnier 1990; Reilly and Seibert 2003); one reason could be the traditional distinction between propositional, denotative, and connotative meaning, and because affect is subjective and often paralinguistic.[8] Perhaps another reason is the difficulty it presents for existing models (Besnier 1990). Below, I describe related work on affective linguistic indicators, resources, and corpus annotation, motivating especially the data-driven approach to affective meaning in text taken in this study.

Automatically predicting affect in text requires knowledge about potential linguistic indicators of affect. Descriptive linguistic attempts illustrate that a wide range of linguistic indicators might signal or carry affect or evaluation in a broad sense (setting aside prosody for Sec. 2.4). In his survey of affect and language, Besnier (1990) described not only emotion words and connotative sense, but also other language phenomena as capable of affective components, with cross-linguistic fluctuations. Examples include (mostly using Besnier's terms directly): terms of address or kinship, diminutive or augmentative morphology, honorifics, (T/V) pronouns, ideophones, sound symbolism, reduplications, intensifiers, quantifiers and comparatives, mood/modality, onomatopoeia, exclamations, insults, interjections, curses, expletives, imprecations, case-markings, and syntactic constructions like left-dislocation, inversion, topicalization/focus, hedges, clefting, word order choices, and raising. An aspect of evaluation in language is distancing, e.g. depersonalization, agentless passives, or pronoun deletion may reflect a distancing perspective. Besnier also mentioned that communication involving quoting, replaying, or code-switching may serve affective or distancing functions, as well as that genres, poetic devices, speech acts, and performance styles might be marked by affect, whereas aspects like emoticons and flames in e-discourse exemplify the importance of affective meaning in writing, and non-lexical behaviors such as laughing, weeping, as well as silence, dysfluencies, stuttering, withdrawing

[8] Also, traditional formal semantics discussing subjectivity tend to be theoretical and detail-oriented, and I do not enter into that literature here.

or being inarticulate, also may indicate affect, as can a conflicting message (e.g. sarcasm or irony). Additional linguistic signals of affect or expressiveness mentioned for example for narratives or in a broader context (Collier 1985; Foolen 1997; Reilly and Seibert 2003) may involve, e.g. intensifiers, modals, hedges, casuals, negatives, affective or mental predicates, certain affixes, various speech errors, certain part-of-speech ratios, level/ratio of negation, referential strategies, theme repetitions, commands, and grammatical complexity, as well as speech acts such as warning, complimenting, thanking, apologizing, condolences, and congratulating.[9] While acknowledging that many linguistic indicators are capable of affective, expressive, and evaluative qualities, and that some of them may be useful for feature set creation for computational methods, a complication is that these signals appear flexible. The presence of one or more of these indicators is not a guarantee for identifying affects. For example, they may perhaps signal multiple affects or other types of subjective language, or they may apply rather at larger text spans, or occur also in non-affective communication, depending on the contextual situation.

As opposed to the current study, much work exploring affective language or meaning seem to have had smaller or qualitative scope.[10] For example, attitudinal language has been studied from discourse analytic, anthropological, cultural, or critical perspectives (Besnier 1990; Niemeier and Dirven 1997; Palmer and Occhi 1999). And, as a particular example, metaphorical phenomena may also involve affective aspects (Besnier 1990), and cognitive linguists have considered emotion categories via the interesting area of conceptual metaphor, e.g. (Taylor 1995; Huang, Yang, and Zhou 2005). Another example includes arguments put forth about the relevance of culture for emotion words and how semantically primitive scripts in imaginary cognitive prototype scenarios can capture emotion terms, e.g. (Wierzbicka 1986) and Wierzbicka's approach discussed by Goddard (1998).

Moreover, several psycholinguistic studies have examined affective lexis, e.g. (Johnson-Laird and Oatley 1989; Whissell 1989; Bradley and Lang 1999).[11] Lexical resources created through such work may be useful for automatic prediction of affect in text. One issue with affective lexical resources from psycholinguistic experiments is sense ambiguity or context-dependent meaning (cf. senses of the word *cry*).

Another important area is text annotation because it creates resources for large-scale quantitative and computational linguistic exploration. Some examples of subjective language annotation and adult commu-

[9] See also the overview on language and emotion by Winton (1990), discussing dimensional models in more details, as well as touching upon emotion models, communication involving emotion, and multimodal studies. Reilly and Seibert (2003) had particularly interesting discussions on sign language, and among other things also summarized studies on emotion words. And for example Andreasen and Pfohl (1976) as well as Fraser, King, Thomas, and Kendell (1986) give examples of linguistic variables examined for psychiatric disorders.

[10] Exceptions include, e.g. a dissertation which undertook a corpus study on the larger issue of stance in English texts via multi-dimensional analysis (Precht 2000), as well as a corpus study of words and expressions' properties of valenced sentiment and evaluative judgement in Swedish newspaper editorials (Hedquist 1978), whose study and linguistic tests rather seemed insightful for opinions, subjectivity, and sentiment.

[11] Whissell's Dictionary of Affect has been used in synthesis for selecting emotional speech units (Hofer, Richmond, and Clark 2005).

nication include Wilson and Wiebe's (2003) work on annotating press using Quirk and colleagues' notion of "private state", Litman and Forbes-Riley's (2004a) discussion on (primarily) polarity labels for annotated dialogue data, and subjective language annotation (e.g. subjectivity and flames) has been discussed by Wiebe and colleagues (2004).[12] Nevertheless, more available resources with affect-specific annotated text data seem needed, as well as annotations extending beyond the adult written domains.

However, at the same time corpus annotation has its down-sides. Annotation studies generally report on interannotator agreement, but as noted by Zaenen (2006) "[f]or interannotator agreement, it suffices that all annotators do the same thing. But even with full annotator agreement it is not sure that the task captures what was originally intended" (p. 577); it should not be confused with understanding a linguistic issue. Nevertheless, corpus annotation may provide resources for linguistic exploration and computational applications. Moreover, agreement scores can be an artifact of annotation scheme and procedure. For example, pairs might be trained to annotate similarly, and across-the-board rules might remove subtle decisions (e.g. questions are negative, or *think* is non-factual), or problematic items might be removed. While these approaches may yield higher kappa and interannotator agreement, cleaner training data, and probably better performance as well as perhaps more consistent applications, the relevance of that for study of linguistic behavior seems less clear. Also, agreement scores reflect the underlying measure, how multiple labels are tackled, etc. Perhaps the concept of a "ground truth" does not necessarily hold for all language phenomena, especially when annotation tasks are highly subjective.[13] The corpus discussion in Ch. 3 applies an interesting diagnostic alternative by looking at the ratios of (dis)agreement types, and avoids the concept of "ground truth" in favor of the notion of "acceptability".[14] Other interesting alternatives to explore disagreement reasons may be G-theory as recently highlighted by Bayerl and Paul (2007), or, e.g. ANOVA studies. Whether the motive is to understand factors underlying disagreements, or as a result make annotations more consistent, e.g. to improve system performance, seems a separate issue.

The initial part of this section suggested that affective meaning is important, and that several linguistic indicators may at times signal aspects of affective meaning. Moreover, it seems that additional corpus-based and quantitative work on the complexities and intricacies of affective meaning and perception in text could benefit linguistics, and more specifically work on texts which go beyond adult writings such as tales. In addition, annotated corpus development may contribute by providing additional resource for large-scale

[12] J. Read has also released a relevant corpus, see `http://www.informatics.sussex.ac.uk/users/jlr24/research.html`

[13] Campbell (2004) noted that "[W]hen producing a database of labels to annotate a corpus, we fall too easily into the trap of thinking that there is one right answer in any situation [...] However, for labelling of affective information in speech, we find that different people are sensitive to different facets of information [...] there is no one-right-answer where the labelling of affect is concerned" (p. 884), and that emotions may co-occur in the speech signal. A representation with probability vectors was suggested.

[14] Traditional agreement scores for the corpus described in Ch. 3 would be low.

exploration of affective language phenomena in written texts.

2.3 Affect and subjective language in text-based NLP

Computational work on subjective language seems to be flourishing and gaining increased attention in NLP. The focus of the following overview is to describe several tasks, systems, automatically developed resources, findings, and application domains, which in one way or another relate to automatically predicting affect in text. The previous work both motivates the task of predicting affect in text itself, as well as the development of an annotated corpus and the approaches taken in the empirical work in this study.

Several types of automatic prediction or classification tasks involve evaluation in text more broadly. For example, tagging *thought units with topic-related attitudes* or *dialogue acts* (Chambers, Tetreault, and Allen 2004), detecting *subjective* sentences, expressions, and "opinion pieces" (i.e. documents representing certain press categories) (Wiebe, Wilson, Bruce, Bell, and Martin 2004), measuring *strength* of opinionated and other subjective clauses (Wilson, Wiebe, and Hwa 2004), or determining adjectival *polarity* (Hatzivassiloglou and McKeown 1997). A particularly popular area is distinguishing reviews' or other texts' attitudinal positive or negative valence in "thumbs-up, thumbs-down" manner through *sentiment analysis*[15] for documents, e.g. (Turney 2002; Pang, Lee, and Vaithyanathan 2002; Dave, Lawrence, and Pennock 2003; Gamon 2004; Read 2005), additionally sentences (Yu and Hatzivassiloglou 2003), or contextualized expressions (words and phrases) (Wilson, Wiebe, and Hoffman 2005), and inferring *attitudinal rating scales* (Pang and Lee 2005). Also, neutrality is complex but important to consider; Wilson, Wiebe, and Hoffman (2005) found sentences with mixed neutral and polarity expressions more common than just mixed polarity, and Koppel and Schler (2006) pointed to the importance of also considering neutral instances for polarity classification. Work on sentiment or appraisal analysis may involve analysis of evaluative lexis, e.g. (Taboada and Grieve 2004; Whitelaw, Garg, and Argamon 2005; Owsley, Sood, and Hammond 2006). Adapting a systemic-functional appraisal taxonomy by Martin and White, which further subdivided attitude into *affect*, *appreciation*, and *judgement*, Whitelaw, Garg, and Argamon (2005) interestingly characterized affect as "the most explicitly subjective type of appraisal". Also, Gordon, Kazemzadeh, Nair, and Petrova (2003) used manual local grammars for finite state transducers in order to recognize expressions of commonsense concepts in text, which was also applied to enhance statistical learning.

Other work has focused on establishing lexical resources which for instance could be helpful for applications and tasks involving evaluative lexical items, such as feature engineering in affective computing. As

[15] Lee (2004) argued that sentiment analysis could be useful for business intelligence, and the ethical implications for its use in government intelligence even received news attention (Lipton 2006).

an example, Esuli and Sebastiani (2006) used classifier techniques to develop *SentiWordNet* with numerical scores for the positive, negative and objective stance of WordNet synsets. They interestingly noted that about one forth of synsets had some degree of non-objectiveness, however only a small set of terms were surely positive or negative,[16] as defined by their method, and they suggested that adjectives and adverbs seemed to carry partially more subjectiveness in language. Also, Strappavara and Valiutti (2004) developed *WordNet-Affect* with affective hierarchical mark-up for WordNet synsets, and Strappavara, Valiutti, and Stock (2006) also augmented emotional synsets with valenced labels, and explored semantic similarity between affective and generic lexis.

Supervised machine learning approaches to automated subjective language tasks require available labeled data sets. The previous section mentioned for example Wiebe and colleagues' efforts to create manually annotated gold standards for subjective text data. Other approaches involve creating data sets through automatic extraction, such as Pang and Lee's (2002) movie review *polarity data set* based on extracted ratings. Automatically created datasets sometimes entail making rough genre assumptions (arguably somewhat questionable at times). For example, Pang and Lee (2004) automatically extracted a *subjectivity data set* of sentence length "snippets" with the assumption that movie reviews were subjective, whereas movie plot summaries were objective. An additional alternative is author-annotated data, e.g. mood labels assigned to weblog posts (Généreux and Evans 2006), however the availability of such data seems domain or genre dependent.

The importance of affect in text for human-computer interaction has been recognized. For example, Karla and Karahalios (2005) discussed an interface for visualizing the expression of emotional tone in text, Mathieu (2005) an interface for navigation by emotional linkage in French texts via an ontology of emotional words, and Subasic and Huettner (2000) visualized the affective tone of texts via "fuzzy semantic typing".

Work specifically on automatic affect prediction in text seems fairly scarce. In terms of methods, for instance Liu, Lieberman, and Selker (2003) mentioned keyword spotting, lexical affinity (associating lexis with affects via probabilities), hand-crafted models, statistical NLP, and their own method, corpus-based conceptual modelling on a large common-sense knowledge base (discussed below). Although data sparsity, domain dependence, and surface feature engineering present crucial challenges to machine learning, learning has been applied to many natural language problems cross-linguistically and in various domains, naturally fit classification, and is arguably related to cognitive inference. Moreover, the boundaries between methods seem blurred and most have drawbacks, some of which span the spectrum. Combining selected approaches seems a plausible framework, and may even present an opportunity to draw on diverse strengths. Similarly,

[16] They noted previous similar observations by other scholars comparing human and automated agreement.

Liu, Lieberman, and Selker (2003) noted that "all of these methods have their merits and can play a role in mutual disambiguation and mutual reinforcement" (p. 127).

A number of approaches and application domains have been considered for predicting affect in text. For instance, Ries and Waibel (2001) classified utterances and segments in meeting data for tripartite emotion. Généreux and Evans (2006) performed classification of weblog posts with author-annotated mood, fit into moods or quadrants in a two-dimensional model of evaluation and activity.[17] Improvements over the baseline depended on specific binary classes. Another study focused on happiness in blogposts and lexical items from blogposts from the same source (Mihalcea and Liu 2006). Among other things, unigrams were used to classify *sad* and *happy* author-annotated blog posts with a substantial improvement in average classification accuracy.[18]

An intuitive application for affective text inference is adding emotional intelligence to chat interfaces. Holzman and Pottenger (2003) performed classification of chat messages annotated for emotions. Avoiding subtle examples and focusing on annotating clear one-sided emotions and considering everything else *neutral*, classifying the largest emotion class *happy* against *neutral* worked better than separating *emotional* from *neutral* (they also classified only various emotive targets).[19] Another system attempted to detect and visualize emotion in chat messages via embodied conversational agents (Ma, Prendinger, and Ishizuka 2005), and a dictionary-based "emotion extraction engine", intended for a chat group interface, also providing intensity (Boucouvalas 2002) was tested on novel chapters and questionnaire responses to emotional situations, reporting on near-perfect ratios of "correctly extracted" sentences (however, that concept and the comparison standard seemed unclear).[20] A later description of their "text-to-emotion system" mentioned taking affective history in combination with intensity into account within what they term "the theory of emotional momentum" (p. 185) to display expressive images based on text analysis (John, Boucouvalas, and Xu 2006).

Email is another application area for affect inference in text. Within the context of an affective email agent, Liu, Lieberman, and Selker (2003) performed sentence-level "affect sensing" using six basic emotions (they also tagged sentences *neutral*, and their interface additionally displayed more decayed emotions). An

[17] If mood-indication was optional, it is possible that less noisy data was involved.

[18] Mihalcea and Liu (2006) also composed a list of happiness-affiliated words and found their *happiness* score weakly correlated with pleasure and dominance in the ANEW word list. Interestingly, several *happy* unigrams reflected, e.g. social behaviors (e.g. *shopping, bought, lunch*, and *drunk*), and ornate descriptions, objects were also common, and high-ranked *happy* bigrams often involved *new* things and the birthday concept. *Sad* words referred to actions and were also more human-centered given WordNet relations and larger n-grams. The temporal distribution in web queries of words inferred varied by time and day of the week. They concluded their paper with an amusing recipe for happiness.

[19] Using different pairs of chatters for training and testing showed similar but somewhat lower results. Holzman and Pottenger's (2003) feature set included among other things phoneme-counts. Whissell (1999, 2000, 2001) analyzed phoneme distributions and found that certain phonemes occurred more promptly with certain emotional affinities, indicating affective semantic properties of phonemes.

[20] Sentences containing emotion words from a tagged dictionary were extracted as emotional, and adjectives modifying emotion words increased intensity. Certain sentences, e.g. starting on auxiliaries, were disregarded as non-emotional.

interesting iterative propagation over a common-sense knowledge text collection, rooted in affect lexicon, was used to develop common-sense based affect models, combined to a representation of the emotion of a sentence, while interpolation and decay factors modelled emotional transitions. Evaluation considered user satisfaction, rather than reporting on system classification performance.

Lastly, text-based affect prediction has to some extent been ignored in affective text-to-speech synthesis. An option to avoid automatic emotional inference is to assume that users select the expression type (Eide, Aaron, Bakis, Hamza, Picheny, and Pitrelli 2004), but a sophisticated system ought to act more independently. Most relevant to this research is work by Francisco and Gervás who attempted automatic prediction of emotion for a storytelling system generating simple versions of fairy tales. Their system EmoTag marked up sentences with scores for three affective dimensions using mostly keyword spotting and average lexical affinity for pleasantness (P), activation (A), and dominance (D), based on word lists with dimensional scores using WordNet extensions. Francisco and Gervás (2006b) determined success by dimensions given a somewhat obscurely described deviation-based procedure, which compared the system against human judges, whereas another work (Francisco and Gervás 2006a) also mentioned categorical emotion labels via keywords and lexical affinity for sentences in tales, where a bit less than half of sentences were marked the same way by a majority of annotators. As one would expect, the success rate was lower for stories not used to create one of the word lists used. They concluded that dimensional tagging was more successful and improved over or equalized human judges, and that the dimensional approach was preferable to apply to speech. However, their human judges felt that categorical labels were easier to assign, which seems important since text-to-speech starts from text analysis; their comparison standard was unclear when they argued that categories and dimensions mapped quite well. In another paper, they also dealt with negation via dependency structures, inverting word list entries' PAD estimations. EmoTag was a laudable attempt to annotate emotion in text for a text-to-speech storytelling system, and they rightly noticed that lexis was important for affect inference. Nevertheless, they had a modest corpus size, and their prominent focus on keyword spotting and lexical affinity seemed to leave room for alternatives, whereas the dimensions as well as the deviation-based evaluation metric complicated comparisons.

The above summary discussed selected related NLP tasks, applications, lexical resources, ways used to develop subjective datasets, methods and applications for affect prediction in text, etc. The summary showed that there exists a clear interest in subjective and affective NLP, motivating the importance of the task pursued in the current study. It also indicated that there may be a need for more sophisticated approaches to predicting affect in text, and an option is combining various methods in joint prediction. In addition, most studies focused on adult communication, whereas child-directed texts have received less attention, and additional work to create available resources that enable computational methods appear needed, especially for classifying sentence-level affect in text. This study moves towards addressing these issues.

2.4 Affect in phonetics and speech technology

The prosody of speech is a key component for conveying affect (Besnier 1990). Shih (forthcoming) provides an overview of issues in prosody as well as emotional prosody, and notes the importance of comprehending the functional, form-related, and variational properties of prosody, e.g. for speech technology implementations.[21] The second part of this dissertation project deals with exploring emotional prosody, and this section describes the complexities of affect in speech, situating the approach taken in this work within the larger context of research in phonetics and speech technology.

Prosody involves various acoustic factors, such as the voice's pitch (or fundamental frequency/F0, perceived as pitch), amplitude, speech duration and tempo, pausing, voice quality (e.g. creaky or breathy voice), formants, spectral tilt, physical jaw gestures, and spectral features. In signed languages, prosody reflects changes in tempo, duration, and shape of signing form (Reilly and Seibert 2003). As discussed by Shih (forthcoming), prosody serves multiple functions in language, which are characterized by cross-linguistic and intra-linguistic variations as well as similarities. The linguistic aspects of prosody increase the challenge in teasing apart paralinguistic or affective prosodic phenomena; for example, stress languages like English, pitch accent or accentual languages like Swedish, or tone languages like Mandarine Chinese also encode meaning distinctions of lexical items with prosodic changes, and declarative sentences versus various question types have different F0 curves, with cross-linguistic distinctions to various degrees.[22]

Shriberg (2005) identified expressiveness as one of the main challenges for speech research. Variability in the signal is important for naturalness (Murray and Arnott 1996), and affective prosody is an important aspect of speech diversity. Describing, identifying, and classifying emotional prosody in human speech has already been dedicated much research, e.g. (Murray and Arnott 1993; Banse and Scherer 1996; Dellaert, Polzin, and Waibel 1996; Roy and Pentland 1996; Polzin and Waibel 1998; Cowie and Cornelius 2003; Scherer 2003; Reilly and Seibert 2003; Devillers, Vasilescu, and Mathon 2003; Zhang, Hasegawa-Johnson, and Levinson 2003; Oudeyer 2003; Hasegawa-Johnson, Levinson, and Zhang 2004; Schuller, Müller, Lang, and Rigoll 2005; Chuenwattanapranithi, Xu, Thipakorn, and Maneewongvatana 2006). To determine prosodic parameters or rules of particular emotions or affective dimensions, scholars might review literature, analyze

[21] Listeners can perceive affect in speech (e.g. towards the content or the audience), and speakers can convey an affective state (Polzin and Waibel 1998); both sides of the coin are important in a communicative situation.

[22] There are many models of intonation (Dutoit 1997; Kochanski and Shih 2003), such as Silverman and colleagues' ToBI, Fujisaki's phrase and accent commands, British *tonetics*, van Santen's intonation model, and Kochanski and Shih's Stem-ML. For prosodic modelling in text-to-speech synthesis, see e.g. (Dutoit 1997; Sproat 1998).

affective speech corpora, or vary prosodic parameters in a systematic fashion (Francisco, Gervás, and Hérvas 2005). Besides naturally occurring speech with discourse content, scholars may explore speech with neutral or nonsense semantics, or speech processed to be meaningless (e.g. by filtering or reversing). Most work on characterizing natural emotional speech has been conducted at the global sentence level or without exploring impact on local linguistic units (Schröder 2001). Less seems known about local effects, e.g. at the word level.[23] What local means depends on one's definition, although aligning it with a linguistic unit (e.g. phrase, word, syllable, phone) seem appropriate for linguistic exploration, compared to low-level time windows. For example, Linnankoski and colleagues (2005) investigated emotional prosody in a single word (*Sarah*), and Rotaru and Litman (2005) reported encouraging results and turn-length effects when exploring word-level pitch for recognizing emotional vs. non-emotional dialogue turns. It seems important to continue empirical exploration of prosodic locality phenomena in different utterances with multiple words.

Much work has gone into analyzing emotions' prosodic profiles, e.g. (Murray and Arnott 1993; Scherer 2003), and some prosodic trends reappear. For example, *anger*[24] is often characterized by increase (in speech rate, F0 mean and range, and intensity), whereas *sadness* is mostly described as marked by opposite behavior. In signed language, *sad* versus *angry* prosody was distinguished by longer and shorter signs and shape movement, and these emotions "exhibited the most consistent differences" (p. 539) (Reilly and Seibert 2003). Scherer (2003) also noted that *sad* and *angry* are often best recognized in the audio modality.

Nevertheless, studies on emotional speech have yielded varied findings, and sometimes contradictions (Scherer 2003), and emotional acoustic trends and their variations still require research. More complex properties have also been found for more natural speech compared to acted speech (Batliner, Fischer, Huber, Spilker, and Nöth 2000). Speaker-dependent effects also occur, e.g. (Roy and Pentland 1996; Schröder 2001). Across languages, for example Abelin and Allwood (2003) examined human classification accuracy for affective categories in cross-linguistic perception with Swedish emotive speech stimuli (based on manual semantic groupings of open-ended emotion term answers). Their data indicated that although natives mostly did better overall, and particular language groups had more difficulties with certain emotions, several emotions (again, especially *angry* and *sad*), although not all, were quite well-classified by speakers of other languages. These types of perception data support that there are both universal or cross-linguistic and language-specific or cultural components to affective speech; also see Shih (forthcoming), who discusses results pointing in the similar direction by Scherer, with above chance agreement for non-natives and better

[23] In contrast, local prosody has been characterized from lexical and syntactic/semantic points of view, e.g. (Dutoit 1997; O'Shaughnessy 2000). For example, word stress changes due part-of-speech distinction (e.g. noun *'recall* vs. verb *re'call*), and contrastive emphasis, focus, new vs. given information, and final lengthening, F0 continuation rise, or distinctions at the micro-phone level occur.

[24] Many studies also make a distinction between *hot* and *cold anger.* As noted above, they may be characterizable as members of a larger emotion family, and it seems possible that one member is perceived as a more prototypical or representative instance.

agreement for natives.[25] Considering cross-linguistic vs. language-specific links between production and perception of affective speech adds an additional layer of complexity.

Not only prosody affects perception and recognition of natural or synthetic emotional speech and positive/negative valence. Verbal or contextual features or non-lexical speech sounds may also play a role, e.g. (Cauldwell 2000; Min Lee, Narayanan, and Pieraccini 2002; Schröder 2003; van Santen, Black, Cohen, Kain, Klabbers, Mishra, de Villiers, and Niu 2003; Litman and Forbes-Riley 2004b; Alm and Sproat 2005b; Liscombe, Richard, and Hakkani-Tür 2005).[26] The role of non-acoustic features in affective speech perception and recognition adds a promising link between the task of predicting affect in text and its use in text-to-speech. Certain domains may be less affected by non-acoustic cues than others, e.g. Shriberg (2005) mentioned a study on telephone dialogue systems. In addition, biophysical cues may also be considered for recognition (Kim, André, Rehm, Vogt, and Wagner 2005).[27]

Similar to natural speech, there is an interest in emotional synthetic speech. Cahn (1990) did groundbreaking work on English emotional prosody in text-to-speech synthesis, obtaining above chance recognition in perceptual experiments with parameters based on inspecting literature. Burkhardt and Sendlmeier (2000) explored emotional synthesis for German in perceptual experiments with various combinations of acoustic parameter settings. Taking an evolutionary approach, Sato (2005) cast expressive prosody as voice conversion using an interactive Genetic Algorithm to optimize prosody of natural and synthetic speech, but did not specifically discuss the problem of interactive evolutionary computation such as user fatigue and flexible or changing user perception or criteria for search targets (Takagi 2001; Llorà, Sastry, Goldberg, Gupta, and Lakshmi 2005). Many have attempted to synthesize or explore emotional speech with formant, unit-selection or concatenative synthesis, and prosodic rules, emotional speech units, or copy resynthesis may be used. For an overview, see Schröder (2001), and examples include e.g. (Murray and Arnott 1996; Murray, Edginton, Campion, and Lynn 2000; Gobl and Ní Chasaide 2003; Iriondo, Alías, Melenchón, and Llorca 2004; Schröder 2004). For examining emotional speech, resynthesis can also be used to modify parameters of neutral speech (Scherer 2003).

Overall, the matter of emotional prosodic profiling appears far from solved, and novel approaches to exploring emotional prosody may shed additional light on this matter. A reoccurring issue for affective speech research is the limited access there is to affective speech data. For instance, as noted by Aharonson

[25] Also see Oudeyer's (2003) discussion on Tickle's result for Japanese and American listeners, and Chuenwattanapranithi, Xu, Thipakorn, and Maneewongvatana (2006) mentioned Thai listeners' perception of *anger* and *joy* in several languages.

[26] Litman and Forbes-Riley (2004b) also noted that automatically speech recognized based text could under certain conditions be more helpful than real text. And Nass, Foehr, Brave, and Somoza (2001) found that content was more liked when the emotion of the voice and content matched, however, interestingly, mismatch between emotion and content was perceived as more credible.

[27] Articulatory information has also been used for classifying emotional speech (Lee, Yildirim, Kazemzadeh, and Narayanan 2005).

and Amir (2006) "[t]he most common problem in this type of study is obtaining a corpus of emotional speech." Speech data bases are expensive to record, and recording actors portraying emotions is often criticized. Other strategies to get natural affective speech data involve eliciting affective speech, or selecting genuinely affective speech material. Whereas genuine emotional speech materials are mostly preferred, they may not only be difficult to obtain, but even suffer from methodological or legal and ethical issues.

To conclude, this dissertation makes an attempt to further explore resynthesized (i.e. computer-modified) emotional speech within the framework of evolutionary search, using the rigor of a robust *active interactive Genetic Algorithm* (Llorà, Sastry, Goldberg, Gupta, and Lakshmi 2005).[28] Compared to for example studies based on speech corpora, it conceptualizes the exploration of emotional speech as evolving emotional prosody in a flexible manner based on inexpensive resources, as well as examines prosody at the local word-level resulting from this process.

2.5 Tales and oral narratives

Since this research was conducted with an interest in automatic storytelling, and components deal with manual annotation and automatic prediction of affect in written tales, this last subsection discusses a few pieces of selected work in relation to narrative phenomena in storytelling as well as paralinguistics or affect in tales from the point of view of text-linguistics, prosody, and story-related technology. However, because the focus of this work is affect, this section does not intend to approach an overview of the vast existing research on narrative.[29]

First, stories have structures, and these seem related to expressiveness. For example, an exciting narrative may involve a "build-up of dramatic tension followed by a release or resolution at the climax of the story" (p. 5) (Anderson and McMaster 1982). William Labov's work on narrative has been influential with respect to oral narrative in every-day human-to-human dialogues, and Labov and Waletzky's *evaluative* component of storytelling reflects the narrator's affective stance to parts of the narrative, and serves a social purpose of engagement and attention maintenance, cf. (Wennerstrom 2001; Reilly and Seibert 2003).[30] More specifically in relation to folk tales, Propp (1968) suggested certain enumerable and linearly ordered functions.[31] Moreover, Reilly and Seibert (2003) report on a previous study which found more evaluative lexis

[28] Genetic Algorithms have been previous applied to language and speech, e.g. (Banse and Scherer 1996; Redford, Chen, and Miikkulainen 1998; Ke, Ogura, and Wang 2003; Alías, Llorà, Iriondo, and Formiga 2003; Oudeyer 2003; Llorà, Alías, Formiga, Sastry, and Goldberg 2005; Sato 2005).

[29] For instance, computational storytelling also relate to Human-Computer Interaction, story generation, and narrative intelligence. These topics extend beyond the scope of this study and are not discussed.

[30] Evaluative aspects such as expressing a point of view may be more prevalent in everyday oral narratives, compared to reading a fairy tale, although this seems related to degree of storyteller's performance freedom.

[31] Propp's work has been influential for for story generation and story grammars, e.g. (Peinado, Gervás, and Díaz-Agudo 2004).

at points before the story conflict and its resolution for adult hearing tellers. Interestingly, they characterized the affective prosody of the spoken narratives as idiosyncratic.[32] Assuming that story structure is partly related to affect trajectories in stories, automatic approaches to predicting affect in text may benefit from incorporating text features which try to capture affect sequencing.

Previous work shows that expressiveness appears to be important in stories. For example, Bralé, Maffiolo, Kanellos, and Moudenc (2005) considered perception of spoken and written sentences from fairy tales and analyzed affect clues such as certain punctuation, lexicon, and syntax, etc. given three basic emotions. Moreover, a semi-automatic system evaluated *tension* in passages, based on a score considering lexis' evaluation and activation dimensions (Anderson and McMaster 1982). Given their implementation and lexical resources, three Beatrix Potter stories claimed as preferred by children showed peaks of tension around the story middle, and the authors' pondered on whether, e.g. Proppian story structure may be used to predict affect patterns. At any rate, negative states appear important for driving the plot. Another early system was BORIS, which interpreted narrative characters' affective states, reactions, and causes, seemingly based on appraisal and arousal mechanisms (Dyer 1987). The mentioned studies testify to the importance of expressiveness in narratives, and it seems appropriate to investigate storytelling while considering affect or related concepts and explore computational approaches for this purpose.

Narrators are obviously important for storytelling. For example, a narrator can mimic several different characters, which may differ sociolinguistically (age, gender, or by social, ethnic or geographical group) as well as in other ways (e.g. degree of seriousness, "voice size", personality, etc.). Also, Wennerstrom (2001) noted among other things that exaggerated prosody tended to be combined with evaluative language cues in oral narratives (e.g. stories in everyday conversation); this observation again seems to provide promise for linking expressive prosody to automatically predicting affect in narrative, and it also raises interesting questions about locality effects is affective speech. In discussing everyday oral stories told by native and Asian non-native speakers of English Wennerstrom also observed that pitch peaks tended to fall on the evaluative function, and especially on quoted speech (in fact more so for non-natives), followed by expressing viewpoint.[33] In other words, considering quoted material in feature engineering for computational approaches for predicting affect in text may be useful. Moreover, quoted speech is certainly important in automatic storytelling systems, and perhaps especially so when synthesizing affective text passages that represent direct speech since these may require heightened expressivity. Zhang, Black, and Sproat (2003) performed

[32] Moreover, referring to an earlier study by Bamberg and Reilly, which examined elicited children's narratives, they claimed that children as narrators go through developmental stages.

[33] Wennerstrom (2001) interpreted exaggerated pitch as a "performance feature - a manifestation of the speaker's emotional involvement with these evaluative points of the text" (p. 227); and she noted that exaggerated prosody may also accompany aspects such as key climactic events.

automatic classification of quote type, character identification, and speaker assignment. Character identification can be especially challenging in children's stories' magical worlds with talking toys, unnamed people, animals, and objects.

In terms of computerized applications, virtual storytelling and storytelling systems, e.g. (Gustafson and Sjölander 2003; Bernsen, Charfuelàn, Corradini, Dybkjær, Hansen, Kiilerich, Kolodnytsky, Kupkin, and Mehta 2004), fall within Human-Computer Interaction. One problem is a need for realistic expectations on automatic storytelling systems. For example, a synthetic voice does not live up to that of a real human (Cabral, Oliveira, Raimundo, and Paiva 2006),[34] and Blom and Beckhaus (2005), while conceptualizing the theoretical underpinnings of a virtual emotive storytelling system, which their discussion mainly relates to tension, noted that automatic emotion recognition is "in its infancy" (p. 26). The question of how to incorporate emotional inference into storytelling systems has been tackled in various ways, e.g. with manually marked-up text input, a knowledge base, or interactive user manipulation (also see the previous discussion on Francisco and Gervás' EmoTag above). For example, Silva, Raimundo, and Paiva (2003) described an interactive emotive storytelling system, based on the idea that good human storytellers are flexible and adapt to the audience. In their multimodal system, drawing on hand-authored stories allowing for non-linear plots, affect was encoded as part of StoryBits, and mood, as well as other properties, could be manually manipulated interactively by the user in the course of the story, and incorporated gradual affects. This design seems interesting for creative and engaging pedagogical purposes, however automatic affect prediction may be needed for systems interacting with users with disorders that impair affective-social development. More broadly, automatic affect prediction efforts seem more closely linked to increasing AI agents' independence, supporting a vision towards AI systems becoming communicative partners. Another study which emotionally modified the text-to-speech component of a storytelling system interestingly found that the majority of emotions were better recognized in a story than in a semantically neutral sentence (Francisco, Gervás, and Hérvas 2005), i.e. contextual information made a difference, adding an encouragement for pursuing automatic affect prediction in story texts. Also, evaluators commented on the difficulty of assigning emotions, and *sad* and *angry* speech worked best whereas *surprise* and *happiness* were problematic.

Other studies have focused on different aspects than affect in expressive storytelling prosody.[35] An interesting study was provided by Theune, Meijs, Heylen, and Ordelman (2006) for Dutch storytelling speech.[36]

[34] Even when paired with a synthetic character, a real voice was perceived more likable, intelligible, appropriately expressive, and credible, compared to a copied synthetic voice, although synthesis quality and character design may also have had an impact.

[35] Also see (Wennerstrom 2001).

[36] They also noted that narrators "mimic characters' voices, produce various 'sound effects', and use prosody to convey and evoke emotions, thus creating an engaging listening experience" (p. 1137), but then avoided emotional and character speech by labeling them "special effects" (p. 1138).

They focused on *global storytelling speaking style* and *suspense* which they divided into at least unexpected *sudden climax* [37] vs. expected *increasing climax*[38], and took a rule-based approach to accordingly modify prosody of a text-to-speech system, also covering "vowel-stretching". They claimed that the global storytelling style would not require inference (and that it might add suspense by itself), and suggested that a story generation system could add suspense mark-up based on the system's available knowledge of story structure and meaning, but also discussed existing story texts as input and noted that "automatically determining which parts of plain text are suspenseful or need extra emphasis is still an unsolved issue" (p. 1140). Theune, Meijs, Heylen, and Ordelman's work was insightful and I have noticed several of their points independently in a small speech corpus of English storytelling speech, such as suspense/excitement or the default expressiveness marking storytelling. Moreover, suspense seems an important phenomenon for expressive storytelling,[39] considering its relationship to story tension. Theune, Faas, Nijholt, and Heylen (2004) also mentioned Proppian structure as a possible resource for further exploring suspense categories. I do not address suspense directly in this dissertation, although I include a simple tension score in my feature set. Automatically learning suspense from text may require a large database of recorded stories annotated for prosodically marked climax locations; such a resource remains a challenge for future scholarship. Nevertheless, the latter authors' notion of unexpected suspense or sudden climax might relate for example to emotions like *surprise* (or negative affective states like *fear*); establishing this would require further analysis, but if that is the case, suspense may be partially capturable or aided by affect prediction in text.

Thus, stories and storytelling seem to involve narrative sequencing or trajectories, as well as multiple phenomena, e.g. tension, suspense, characters, quotation, affect, cognitive states, etc., and perceptions and behaviors of audience and narrator. This dissertation addresses one such component by looking at affect in text and speech. More specifically in relation to storytelling, this study involves data development and exploration of computational techniques for predicting affect in popular written stories.

2.6 Concluding summary

This chapter has surveyed approaches to characterizing emotion, previous work on affective language in linguistics, subjective and affective language in NLP, emotional prosody in phonetics and speech technology, and some findings on tales and storytelling technology. Each subsection motivated the course of research

[37] Involving steep pitch and intensity increase on introductory keywords such as *then*, *suddenly*, *but* for the dramatic occurrence.

[38] With step-wise increased intensity and pitch when closing in on an expected climax, and a pause before the dramatic occurrence.

[39] I am also grateful to Esther Klabbers and Valérie Maffiolo for discussing the importance of suspense within the storytelling context in conversations. Also, in an informal conversation, William Labov commented on the aspect of interest for affect.

taken in the following chapters. Next, I move on to discuss the annotation project and the annotated corpus collected.

Chapter 3

Corpus data for affective text prediction

This section covers the annotation project. The main purpose of creating a corpus of texts annotated for affect was to have a labeled data set for evaluation comparison and for supervised classification in affective text prediction (see Ch. 4) since no large-scale appropriate corpus seemed applicable, especially given the interest in child-directed storytelling. Moreover, such a data set may also be insightful for examining affect perception in text or for linguistic text analysis. First, I give an overview of the annotation procedure. Next, I survey the labeling scheme used as well as extensions which were made with automatic inference of an affect hierarchy. Lastly, I discuss an interesting subset of annotated affective sentences marked by high agreement and hypothesize about its textual properties.

3.1 Annotation of affect in text and experimental adjustments

Within the overall context of a storytelling text-to-speech application as discussed in Ch. 1, the corpus project targeted the annotation of emotive contents at the sentence-level in three sets of children's stories: Beatrix Potter, H. C. Andersen, and the Brothers Grimm.[1] The six annotators who contributed to the final corpus were female US English native speakers, recruited either from a course on Grimm's Fairy Tales or from library science, and paid hourly for the annotations. They received a relatively short hands-on training, where they practiced using the annotation tool on a brief story while discussing for instance alternative annotation options with the experimenter,[2] as well as a manual (see appendix A) and then worked freely, annotating a sentence at a time in chronological story order. Annotating texts is overall challenging for humans, and this case especially so since affect perception is highly subjective and clear definitions are lacking, as noted in Ch. 2. Thus, the resulting annotations may have been subject to the uncontrolled effects of fatigue, boredom, subjective interpretation, misunderstanding, diverged diligence, desensitization,

[1] Project Gutenberg has works by these authors `http://www.gutenberg.org/` (some modified versions), and recently released audio books for texts by these authors. Thanks to Alan Black for making the H.C. Andersen texts available; a likely source is Paull's English translation from 1872, cf. `http://hca.gilead.org.il/` and `http://hca.gilead.org.il/vt/`.

[2] The training sessions were mainly done individually. Naturally, one cannot rule out that the training discussion or examples or phrases (e.g. *neutral descriptive* or *happy end*) from those conversations affected annotators' future perception, just as one cannot rule out that annotators might have been influenced in some other way. Also, the manual was updated and corrected somewhat during the course of the annotation project (e.g. bullet 7e incorrectly said "primary mood" in earlier version).

uncaught human errors, computer failure causing reannotation, and minor interface or procedural design flaws.[3] Also, extratextual factors such as annotator personality, personal task predispositions, or daily mood fluctuations may have contributed to variation in the corpus data.[4]

The sentence-level was marked for affect contents, and focused mostly on perception of text-internal affect.[5] An alternative would be to estimate extratextual affect responses, e.g. feelings of an audience listening to a story being read. However, within a text-to-speech context, the focus naturally falls on the reader. Similarly to an automatic system, a human who reads a story aloud would receive text-based input, and additionally use background knowledge and experience (and perhaps visual story material) to interpret the emotive and other meaning components in the text, e.g. resulting in telling the story with appropriate affective tone.

Each sentence in the corpus used in the classification experiments in Ch. 4 had **four affect labels** stemming from each of two annotators obligatorily assigning a *primary emotion*[6] and a *mood* to the sentence.[7] Specifying an affective primary emotion required assigning a *feeler* of the emotion, e.g. usually a salient character speaking or acting in the context. NEUTRAL was naturally *feeler*-less. However, in a speech application, simultaneous voices are generally not used. Moreover, it seemed difficult to select a particular kind of label (primary emotion or mood) as being a more appropriate representation throughout the corpus for predicting affect; their distinction is fuzzy, and mood seems especially tricky and was a loose concept (e.g. meant as general mood or background affect of a section of the story). Thus, the primary emotion and mood annotations were combined into one set of affect labels for each sentence based on the idea

[3] A few peculiarities in the GUI or data preparation possibly resulted in some inconsistencies, e.g. due to corrections, changing annotations when correcting, minor visual peculiarities (e.g. a word partially falling outside, or entries appearing duplicated when resizing window), accidental click of primary emotion's button, cursor position, constraint on feeler entry recording, or long sentences appearing shrunk due to computer-dependent issues for certain window resizing or due to extreme length (GUI lacked scroll bars). A minority of Grimm's stories lacked a mood variable when unannotated, but regular mood annotations were produced with the annotation tool.

[4] For example, Picard (1997) commented on that mood of subjects may have an influence. It also seems that saturation and desensitization effects might appear.

[5] Emotion is often assumed to have shorter span, which speaks for a shorter linguistic unit like the sentence; although most later empirical work combined primary emotion and mood because the distinction appeared slippery.

[6] Note that I use primary and secondary in the intuitive sense of more vs. less evident, rather than as meaning more neurological primordial vs. cognitive, as has been done elsewhere, see (Picard 1997).

[7] Requiring more thought-through decisions might reduce boredom and focus the task, and other emotional information marked by annotators included *feeler*, *emotional/connotative words and phrases contributing to emotion assignment* (most likely with focus on primary emotion), or emotional *intensities* (i.e. for the primary emotion, and potentially also indicating *secondary emotions*; a NEUTRAL intensity like 0 or 2 is meaningless). The *feeler* requirement was mainly a strategy to focus the decision-making. Annotators had rough character lists for most stories, with typical types (e.g. "villain", "heroine") or particular character names. High-quality character sets would have required deep reading of each story, which was infeasible, thus ambiguities and incompleteness probably occurred. However, *other* or *reader* worked as *feeler* options when other characters did not apply, and for at least two long stories individual character sets were improvised. The interface also had a problem flag but its usage was explicitly discouraged due to the task difficulty and it was ignored when extracting the final corpus. Beyond the 100th word in a sentence (extremely rare), marked emotion/connotative words were not extracted for these other annotations. I do not cover these other annotations, with the exception of including feeler and words/phrases in examples of high agreement sentences, because I am unsure about their usage across annotators, they involved some design flaws, plus they were not used as subsequent machine learning targets. Also, some mistakes remained, as well as a few manual postcorrections, and a handful of updates ignored for compatibility reasons.

Emotion class
ANGRY
DISGUSTED
FEARFUL
HAPPY
NEUTRAL
SAD
POS. SURPRISED
NEG. SURPRISED

Table 3.1: Non-merged affect categories used for annotating primary emotion and mood

that because affect perception is subjective, the concept of "ground truth" does not really hold, and instead rather any one of multiple labels is just as *acceptable*.[8] Also, joining mood and primary emotion mark-up may equalize "conservative" annotators marking more NEUTRAL, which could be preferable given the storytelling application and this genre's default expressiveness. Moreover, as described below, for experimental purposes, the original eight basic labels were later merged to six basic ones, and NEUTRAL undersampling was used, to avoid NEUTRAL overtraining and prefer affective as opposed to NEUTRAL mark-up in cases of sentences with mixed neutral-affective labels. The early pilot work, mentioned in Sec. 4.3, took place before, and instead considered coarse-grained labels based on tie-broken primary emotion labels on a small preliminary subset. However, tie-breaking was later avoided since it seemed unnatural given the subjectivity of this task, expensive in terms of resources, and would seem to emphasize the notion of a "ground truth", which arguably seems invalid in this subjective context, rather than accepting that different perceptual views can be *acceptable*.

3.1.1 Labeling scheme and affect hierarchy

The eight affect labels used for the corpus mark-up are in Table 3.1 and were mainly inspired by the notion of categorical basic emotions. Using a forced choice was justified by the computational application. As noted in Ch. 2, studies with open-ended labels tend to resort to combining labels into a manageable set of learnable categories. Avoiding labels such as *other* or *N/A* might also encourage decisions on subtle cases.

Ch. 2 showed that agreement on emotion representation is lacking. Thus, the choice of affect labels seems rather application-driven. The arousal or activity dimension seems important for speech. Nevertheless, text-to-speech synthesis originates in text and text analysis, the context and focus of the present task. Cognitive evidence for linguistic categories seem to support the use of categorical emotions, and as noted in Ch. 2, Ekman and colleagues have evidence supporting the existence of certain emotion categories across diverse cultures from facial recognition experiments. Moreover, several empirical studies show above chance

[8] This approach may also indirectly capture some aspects of emotion blends.

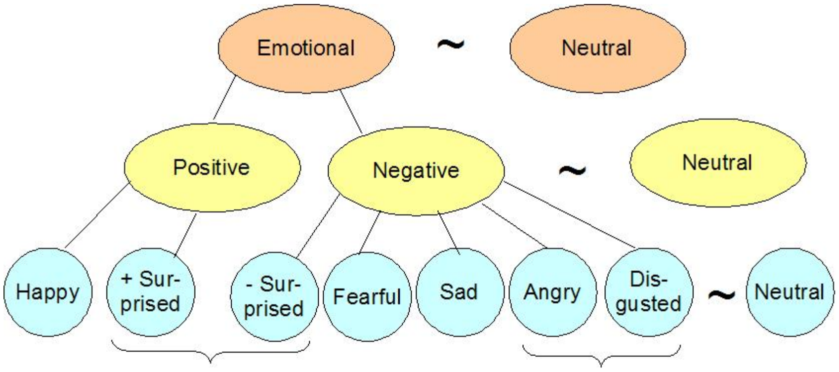


Figure 3.1: Three-level hierarchy of affect labels

performance for recognition of categorical emotions in human and automatic classification tasks involving prosody. Categorical labels may also be more straightforward for annotators to conceptualize compared to dimensional scales, as participants pointed out in the study by Francisco and Gervás (2006b). Also, categorical emotions may be a preferred representation for pedagogy[9] and they naturally fit a computational classification approach. In addition, a basic affect category is broad enough to span a whole subset of related affect states, e.g. the emotion family of ANGRY could also be interpreted as covering concepts such as *irritated, annoyed* and *enraged.* Finally, the foundational nature of these affects intuitively seems to fit fairy tales' contents which may include certain canonical topics and behaviors, compared to more spontaneous natural discourse, although tales seem to also encompass added emotional complexity. Combining categories with intensity also seems a plausible future option for adding an approximation of an arousal dimension, whose relevance for speech was discussed in Ch. 2, whereas in possible future multimodal expansion (e.g. combining text, audio, and vision with a talking head) other modalities may benefit additionally from categorical information.

Given the basic label set, a hierarchical treatment of affect could then be used in subsequent experimentation for automatic prediction of affect, exploring three levels of granularity. The *mid* and *top* levels were inferred from the basic labels, as outlined in Fig. 3.1. The merged *all* level covered six categories: the original FEARFUL, HAPPY, SAD, NEUTRAL, as well as a joint merged class ANGRY-DISGUSTED and the combined class SURPRISED, merging positive and negative cases of SURPRISED. The merges at the basic level were motivated by data sparsity and by the merged categories having related semantics.[10] The *mid* level combined basic labels into three polarity classes: POSITIVE affects (i.e. HAPPY and POS. SURPRISED),

[9] In reference to various studies, Reilly and Seibert (2003) noted that four year old children have quite sophisticated cognitive emotional perception and production. Within the context of discussing development factors they mentioned that "as early as 18-20 months of age children begin to use internal state words and signs, (e.g. mad, happy, or sad) to refer to their own emotional state [...] by [age] 3;6, they can distinguish and label mad, happy, and sad facial expressions. Scared and surprised are correctly identified by the majority of children by four [...] and they can attribute these emotions to others.".

[10] Also, it has been suggested that DISGUST may be more recognized visually than acoustically (Scherer 2003).

NEGATIVE affects (i.e. ANGRY, DISGUSTED, FEARFUL, SAD, and NEG. SURPRISED) and NEUTRAL, whereas the *top* level considered two inferred classes: non-emotional NEUTRAL and EMOTIONAL (i.e. all affective labels). In addition, as mentioned above and in Ch. 4, for most classification experiments, the NEUTRAL labels were undersampled, i.e. removed, from mixed neutral-affective sentences. Also, the hierarchy allowed the *mid* level to address the valence dimension, although in a simplified discretized fashion. Naturally, the hierarchical model could also be extended to involve probabilistic affinity as well as emotional intensity.

3.1.2 Corpus characteristics

All in all, 176 stories annotated by pairs were included in the final corpus.[11] The details of the final three corpus parts are in Table 3.2, showing that more stories were annotated for Grimm's and H.C. Andersen. Moreover, as shown in Table 3.3, some annotators covered more texts.

Subcorpus	#Stories	#Sentences	Mean # sentences (stdev.)
B. Potter	19	1946	102 (78)
Grimm's	80	5360	67 (37)
H.C. Andersen	77	7996	104 (75)
Total	176	15302	87 (63)

Table 3.2: General composition of the three subcorpora

Annotator	B. Potter	Grimm's	H.C. Andersen
AR	19	39	69
SS	-	32	36
AS	19	19	9
LC	-	22	8
RE	-	20	-
KS	-	28	32
Total	38	160	154
Unique (by 2)	19	80	77

Table 3.3: Annotators' count of stories, given the corpus with two sets of annotations per story

[11] The annotation data for the primary emotion and mood labels for two annotators per sentence (before merging ANGRY-DISGUSTED and SURPRISED classes, or doing neutral undersampling) is available as-is (i.e. affect label annotations with sentences as annotated). Sentence split was automatic (treatment of source texts' end-of-line hyphenation differed). Besides some preprocessing (e.g. special whitespace like song indentation was not maintained, and some dedications were removed), some directories have files with the additional suffix/infix *okpuncs*, meaning the sentences were additionally processed (e.g. before POS tagging of sentences) to some degree in a subcorpus-dependent manner but to some degree developed more on Potter (completing double quotes at the sentence level - multisentence quote spans were not considered, removing certain edge punctuation or white space, modifying thought dashes, and revising a quote-apostrophe mixup in Grimm's from preprocessing and split hyphenated words in Potter). This was done with heuristic rules so some inconsistencies remained or were introduced. Similarly, certain POS tags are not satisfactory. Also, some peculiarities of the original texts were not adjusted (e.g. quotation usage in H.C. Andersen). Note as caution that I could not inspect most annotations or other processed files manually due to corpus size. One Grimm's story was split into its two natural parts. Stories lacking a second annotation were excluded from the final corpus, as were a few existing third annotations, and one set was reannotated. The corpus also excluded a training story, and a story missed in the external data processing, as well as annotated story title and THE END sentences in H.C. Andersen.

Fig. 3.2[12] shows the overall ratios of different annotation types for the three different subcorpora, considering the merged six basic classes (the *all* level) and the four affective labels of each sentence, according to Equation 3.1.

$$\text{Overall ratio of annot. type}_i = \frac{\sum_{stories} \text{sentences with annot. type}_i \text{ in story}}{\sum_{stories} \text{sentences in story}} = \frac{\text{total \# sentences with annot. type}_i}{\text{total \# sentences}} \tag{3.1}$$

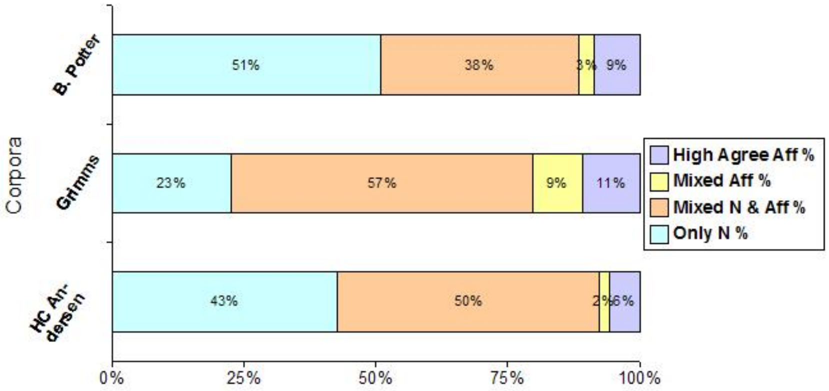


Figure 3.2: Overall ratios of annotation types for the three subcorpora, using 6 affect labels (5 for high agree affect)

At first observation, sentences only marked NEUTRAL were quite common, and more so for B. Potter or H.C. Andersen than for Grimm's. Whereas this could depend on annotator identity and preferences, one could also hypothesize that Grimm's represents more prototypical fairy tales with more affect.[13] In addition, disagreements were a characteristic feature of the affect data, and was consistently more common for sentences marked both with NEUTRAL and one or more affective classes, than for sentences just marked with different affective labels; this seems to parallel findings for polarity expressions in subjective texts (Wilson, Wiebe, and Hoffman 2005), mentioned in Ch. 2. It also indicates that the border between affective and neutral is fuzzy. As mentioned in Ch. 2, it appears likely that since affective perception lacks clear definitions and is subjective, NEUTRALITY (i.e. the absence of affect) also suffers from the same problem. Moreover, it seems some annotators differed in "conservativeness", i.e. annotating more vs. less neutrality. Overall, affective targets appeared to be interpreted and perceived dynamically. Thus, there are bounds on the agreement that can be expected, given the task's high degree of subjectivity.[14] However, interestingly,

[12] View/print this graph in color.

[13] Although one should probably consider that the Brothers Grimm also performed editing, and the faithfulness claim has been critically reassessed, cf. `http://en.wikipedia.org/wiki/Brothers_Grimm`.

[14] Interannotator agreement on a subset of the data for primary emotions was lower than for more straightforward NLP tasks. Similarly, intra-annotator data on a small set indicated that primary emotion targets were dynamic.

sentences marked with only affective labels were more commonly characterized by *high agreement*, i.e. having the same affective class annotated for all four obligatory affect labels for a sentence (given the merged label set at the *all* level), than mixed affective labels. This especially held for stories by B. Potter and H.C. Andersen. In Sec. 3.2, I examine this interesting subset of high agreement affective sentences in more detail for the H.C. Andersen subcorpus.

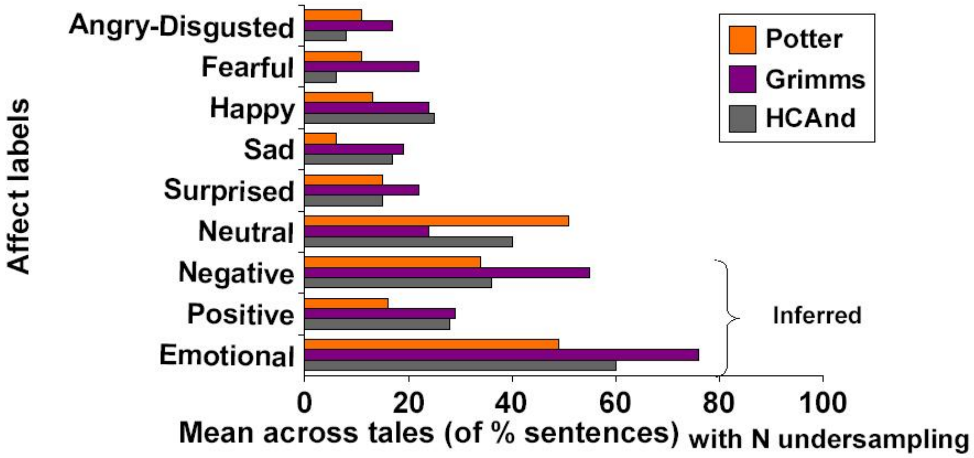


Figure 3.3: Distribution of affect labels as means of % sentences in tales (after ignoring NEUTRAL labels for mixed sentence annotations)

$$\text{Mean of \% sentences with label}_j = \frac{\sum_{stories} \frac{\text{\# sentences with label}_j\text{in story}}{\text{\# sentences in story}}}{\text{\# stories}} \tag{3.2}$$

Additionally, Fig. 3.3[15] compares the means of % sentences with various labels at different levels of the hierarchy across the corpora, according to Equation 3.2, after undersampling the NEUTRAL labels by removing them from any mixed neutral-affective sentences. Fig. 3.3 reveals that the subcorpora had different affective make-up. For instance FEARFUL was less common than other affective labels for H.C. Andersen, but not for B. Potter. These differences could, for example, be related to topical contents, maturity level, and so on. In general, these stories were annotated with more negative than positive affect labels. As noted in Ch. 2, negative states may help driving the plot forward; as story material negative happenings may make more interesting reading than positive ones.

To summarize, this section introduced and motivated the annotation scheme and procedures, as well as discussed the resulting corpus, corpus makeup, trends for annotator disagreements, the affect hierarchy, and adjustments for empirical work. The next section proceeds to discuss a case study of the high agreement subset of the H.C. Andersen subcorpus.

[15] Graph is best viewed and printed in color.

3.2 Interesting subsets of affective high agreement

I will now discuss the high agreement subsets, with a particular analysis of the H.C. Andersen subset of high agreement. As mentioned above, each sentence had four affect labels since each of two annotators assigned both a primary emotion and a mood, and these four labels were then combined into the sentence's affect labels. A *sentence with high agreement* affect was defined as all four primary emotion and mood labels having the same *affective* label, given the merged basic label set (the *all* level) consisting of six labels.[16] Since the focus was on affective sentences, sentences with four NEUTRAL labels were not considered. Table 3.4 shows the size of the subsets of high agreement sentences for the three subcorpora. Grimm's had slightly higher ratio of high agreement sentences, whereas H.C. Andersen's had less. I can only speculate as to why; it may be because Grimm's source was the oral folktale canon, and thus stories may be less marked by personalized style of a particular author. That H.C. Andersen had a lower ratio of high agreement sentences could perhaps be attributed to added story complexity and subtlety.

Author	#Sentences	% of subcorpus
B. Potter	167	9%
Grimm's	580	11%
H.C. Andersen	460	6%

Table 3.4: Data subsets of affective high agreement sentences for the three subcorpora

Although affect words (e.g. *happy* or *sad*) are frequent among high agreement sentences, such words are not always present in high agreement sentences. For example, some sentences below from Grimm's appear to indicate that descriptions of affect-related behaviors such as *hearing noises* or *fainting* distinguished a particular affect. Thus, more than just affect words contributed to sentences being characterized by high agreement for annotated labels. Next, I discuss characteristics of high agreement sentences in more detail.

1. ANGRY-DISGUSTED: *So she cried out, "The king"s daughter shall, in her fifteenth year, be wounded by a spindle, and fall down dead."* (from Briar Rose)

2. FEARFUL: *Then the youngest daughter said again, "I am sure all is not right – did not you hear that noise?* (from The Twelve Dancing Princesses)

3. HAPPY: *But Red-Cap went joyously home, and no one ever did anything to harm her again.* (from Little Red Riding Hood/Little Red-Cap)

4. SAD: *"Now," thought she to herself, "no aid of man can be of use to me."* (from Lily and the Lion)

5. SURPRISED: *When the princesses saw him they fainted.* (from The Gnome)

[16] Given the surprised inference mapping in the affect hierarchy, a few *high agreement* sentences had 2 labels at the *mid* level.

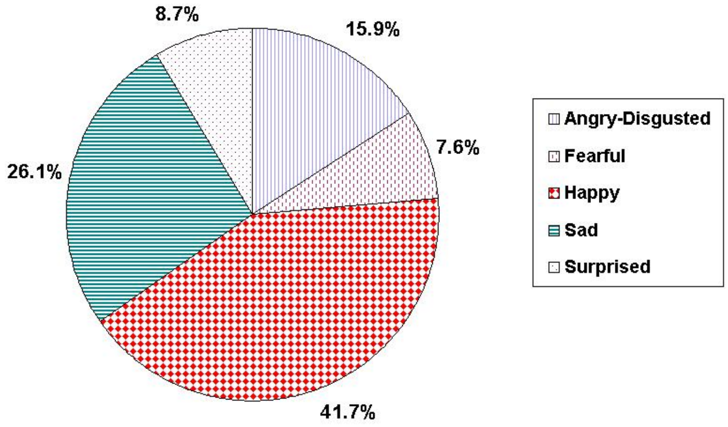


Figure 3.4: Distribution of H.C. Andersen high agreement affective sentences across affective labels (460 sentences)

3.2.1 Interpretive analysis of H.C. Andersen high agreement

This part examines the subset of high agreement sentences in the H.C. Andersen data from a qualitative-interpretive perspective. I dedicated myself to the H.C. Andersen subset (rather than all three subcorporas' high agreement subsets) for feasibility reasons because manual inspection is time-consuming. This analysis is not intended as a rigid categorization attempt, but rather to get an overall idea of why high agreement on both primary emotion and mood across annotators might occur. Also, the coverage of a particular word or phrase in examples does not imply that these necessarily accompany high-agree affect, nor that all or some annotators reacted on these tokens.

This qualitative interpretation draws on my own analysis from examining this subset of sentences, rather than others' detailed mark-up,[17] although I then also looked at annotators' noted *feeler* (see description above) and emotional/connotative phrases for the sentences.[18]

I extracted isolated sentences and mostly examined them that way, rarely considering surrounding context. Naturally, a thorough reading of the involved H.C. Andersen stories could provide alternative interpretations, however such an undertaking would in itself deserve a separate thesis project.

Five annotators were involved in the overall H.C. Andersen subcorpus of 77 tales. 460 sentences were marked by affective high agreement, given the five affective classes at the *all* level. The distribution of

[17] Another alternative could have been to base the analysis on published grammars' characterization of emotion/expressiveness, which might provide an interesting comparison.

[18] Although the use across annotators of these other annotations (*feeler* and emotional/connotative phrases) was unclear, and there were some interface and procedural design flaws, this additional information appears valuable within this qualitative discussion. On the examined subset, often annotators agreed on the *feeler*, whereas emotional/connotative words or phrases often partially overlapped, although cases of absolute or no overlap as well as empty mark-up occurred; some people might have preferred longer vs. shorter phrases. Wilson and Wiebe (2003) found that spans and boundaries of expressions often differed between judges.

affective classes for this subset is in Fig. 3.4, with HAPPY and SAD being most frequent.[19]

Overall, it appears that some sentence characteristics might have assisted the high agreement on affect perception for this subset. I discuss and illustrate several such characteristics below with selected examples: *affect words*, *words for related or contrastive affective states*, *affect related words or expressions*, *polarity words and expressions*, *acquired knowledge and human experience*, *speech acts*, *types of direct speech*, and *mixed emotions*.[20] Naturally, these characteristics occur in some and not all sentences; some appear frequently, whereas others are more rare. Oftentimes, several characteristics may together characterize a sentence, as shown in some of the below examples.[21]

For the sentence examples in this section, affect labels are in small caps, sentences in italics, phrases in bold-face illustrate my point for the particular discussed characteristic, whereas I underline annotators' noted phrases (single underscore for non-overlapping vs. double underscore for overlapping mark-up), and include their *feeler* (with annotator subscripts to show if they had indicated the same or not) in parenthesis in small caps. For instance, in the first sentence example below for *Affect words* the high agreement affective label ANGRY-DISGUSTED is followed by the actual sentence in italics. In the sentence, the phrase ***angrily*** is marked in boldface to highlight that this phrase illustrates the discussed characteristic, whereas the parenthesis following the sentence show that both annotators had VILLAIN as *feeler*, and the underlined phrases show that one of them marked the phrase *he drew his sword and brandished it,* whereas both had marked the word *angrily*. Punctuation was not tokenized for sentences presented to annotators.

- **Affect words.** Verbs, nouns, adjectives and adverbs that directly name an affective state (e.g. reflecting a particular intensity) are common in high agreement sentences. That narration can directly announce affective states is an indication of the important narrative role affect can play in stories. Also, Wilson and Wiebe (2003) noted that annotators agreed more strongly with strong subjective expressions, to which these cases belong.

  Examples:

  1. ANGRY-DISGUSTED: *anger*, *angry*, *angrily*, *contempt*, *cross*, *mad*, *vex*

  2. FEARFUL: *afraid*, *fear*, *fearful*, *frighten*, *frightened*, *horror*, *terrified*, *terror*

  3. HAPPY: *blissfully*, *cheer*, *contented*, *delight*, *delighted*, *delightful*, *gay*, *glad*, *gladden*, *gladness*,

[19] HAPPY and SAD were also more frequent in Grimm's high agreement, and for Potter FEARFUL and ANGRY-DISGUSTED were.

[20] Wilson and Wiebe (2003) mentioned a few similar phenomena with other terminology, e.g. their "expressive subjective elements" approaches *polarity words and expressions*, "speech event" *direct speech*, and "private state actions" *affect related verbs*. Also, Strappavara, Valiutti, and Stock (2006), inspired by Ortony, Clore and Foss, used the terms *direct* vs. *indirect* affect words.

[21] Co-occurring language cues were also noted in the analysis of fairy tale sentences with three basic emotions by Bralé, Maffiolo, Kanellos, and Moudenc (2005).

happy, *happily*, *happiness*, *jovial*, *joy*, *joyfully*, *joyous*, *merry*, *merrily*, *overjoyed*, *rejoice*, *rejoicing*, *satisfied*

4. SAD: *agony*, *despair*, *distress*, *gloomy*, *grief*, *grieve*, *grieved*, *love-sorrows*, *low(-)spirited*, *melancholy*, *miserable*, *misery*, *mourn*, *mournful*, *mournfully*, *mourning*, *sad*, *sadly*, *sorrow*, *sorrowful*, *sorrowfully*, *sorrowing*, *sorry*, *unhappy*
5. SURPRISED: *alarmed*, *astonished*, *astonishment*, *shocked*, *shocking*, *startled*, *surprised*

Sentence examples:

1. ANGRY-DISGUSTED: *They buzzed round the prince and stung his face and hands;* ***angrily*** *he drew his sword and brandished it, but he only touched the air and did not hit the gnats.* (VILLAIN$_{1,2}$)
2. FEARFUL: *A fiery dragon could not have* ***frightened*** *the little boy so much as did the little dog in this place.* (HERO$_{1,2}$)
3. HAPPY: *"Now I am in the house of God," she said, "and in that house we are* ***happy."*** (MAIN CHARACTER$_{1,2}$)
4. SAD: *The sisters* ***mourned*** *as young hearts can* ***mourn,*** *and were especially* ***grieved*** *at the sight of their parents'* ***sorrow.*** ((TRUE) SISTERS$_{1,2}$)
5. SURPRISED: *Great quantities of fish could be seen through the clear water, sparkling in the sun's rays; they were quite* ***surprised*** *when they came so suddenly upon such an unexpected sight.* (HERO$_{1,2}$, HEROINE$_{1,2}$)

Additional special cases: These include **negation** (e.g. *not happy* for SAD); **figurative or idiomatic phrases** (e.g. *one of his heartstrings had broken* for SAD; *overflowing hearts* for HAPPY; or *fly into a passion* for ANGRY-DISGUSTED); affect words occurring with **more than one affect** (e.g. *anguish* for SAD OR FEARFUL); and cases with **related or mixed affect words**, exemplified below.

- **Words for related or contrastive affective states**. Expressions naming related or contrastive affective states in the sentential context may also help evoke a particular affect.[22]

  Examples: *dull, pride, proud, proudly, relief, shame*

  Sentence examples:

  1. ANGRY-DISGUSTED: *It must be done, for the very* ***shame;*** *you are sure to be asked when you come back if you have seen everything, and will most likely be told that you've omitted to see what was best worth seeing of all.* (WIDOW$_{1,2}$)

[22] Naturally, one must also consider that these affective states were not part of the label set made available to the annotator.

2. ANGRY-DISGUSTED: *But the buckwheat spread itself out with* ***pride****, and said, "Stupid tree; he is so old that grass grows out of his body."* (BUCKWHEAT$_{1,2}$)

3. HAPPY: *And there she sat as* ***proudly*** *as if she were in a state coach, and looked all around her.* (MAIN CHARACTER$_{1,2}$)

4. HAPPY: *They looked at Little Claus ploughing with his five horses, and he was so* ***proud*** *that he smacked his whip, and said, "Gee-up, my five horses."* (HERO$_{1,2}$)

5. SAD: *Oh, what bitter tears she shed! and she could not open her heart to any one for* ***relief****.* (HEROINE$_{1,2}$)

6. SAD: *Ida and Anna Dorothea wept bitterly, Joanna stood, pale and* ***proud****, biting her lips till the blood came; but what could that avail?* (ANNA$_{1,2}$, IDA$_{1,2}$)

7. SAD: *"I cannot endure it," said the tin soldier, who stood on a shelf, "it is so lonely and* ***dull*** *here.* (TIN SOLDIER$_{1,2}$)

- **Affect related words or expressions.** Lexical items and phrases which describe actions, properties, behaviors, cognitive states, or concrete objects associated with particular affects occur frequently in the examined high agreement subset.

  Examples: *amused*, *amusing*, *amusement*, *banished*, *beat*, *brandish*, *beg*, *bewilder*, *bewildered*, *care*, *caress*, *clap hands*, *condescend*, *cry* (=weep), *dare*, *despise*, *dreadful*, *droll*, *embrace*, *enjoy*, *enjoyment*, *excited*, *excitement*, *faint*, *fainting*, *faintness*, *feel*, *feeling*, *funny*, *gratified*, *groan*, *growl*, *grumble*, *hatred*,[23] *hide*, *insult*, *jokes*, *kiss*, *laugh*, *laugh at*, *lament*, *like*, *love*, *lovingly*, *mock*, *nod*, *offended*, *oppose*, *pain*, *painful*, *pleasant*, *pleased*, *pleasing*, *pleasure*, *rude*, *scold*, *shake head*, *shudder*, *sigh*, *smile*, *smiling*, *smilingly*, *sob*, *sneer*, *spit at*, *stare*, *sting*, *suddenly*, *suffer*, *tear*, *tearful*, *tremble*, *trembling*, *troubled*, *unexpected*, *venture*, *weep*, *whisper*, *whistle*, *wipe one's eyes*, *wretched*

  Sentence examples:

  1. ANGRY-DISGUSTED: *Her Puggie had seated itself on the ground while she wrote, and* ***growled****; for the dog had come with her for* ***amusement*** *and for the sake its health; and then the bare floor ought not to be offered to a visitor.* (PUGGIE$_{1,2}$)

  2. ANGRY-DISGUSTED: *"It must be a kind of garden plan," said another; and so they* ***sneered*** *and* ***despised*** *the plant as a thing from a garden.* ((OTHER) PLANTS$_{1,2}$)

  3. FEARFUL: *How the fir-tree* ***trembled!*** (MAIN CHARACTER$_{1,2}$)

[23] Here, *love* and *hatred* are regarded cognitive states rather than emotions.

4. FEARFUL: *At these words an icy **shudder** ran through the sparrow-mother.* (SPARROW MOTHER$_{1,2}$)
5. HAPPY: *They **laughed** and they **wept**; and Peter **embraced** the old Fire-drum.* (HERO$_1$, (TRUE) MOTHER$_2$, (TRUE) FATHER$_2$)
6. HAPPY: *"I **like** that!" said the Princess.* (PRINCESS$_{1,2}$)
7. SAD: *The lieutenant **felt** this very keenly, and therefore leaned his head against the window-frame, and **sighed** deeply.* (LIEUTENANT$_{1,2}$)
8. SAD: *With silent steps, still **sobbing**, they left the room.* (FAMILY$_{1,2}$)
9. SURPRISED: *It **bewildered** him as he thought of it.* (COUNSELLOR$_{1,2}$)

Some of the more prominent affect related lexical items include *weep*, *kiss*, *laugh*, *cry* (= *weep*), and various forms of *pleasure*,[24] *tears*, and *smile*. Affect related expressions may **not necessarily occur with just one affect**. For example, expressions of weeping or tears are frequently associated with sadness, but may also depict happiness, and **vividly visualize** a character's physical state and appearance (e.g. *tears running down cheeks*, *cry bitterly* etc.).[25] Naturally, **negations** also occur (e.g. *no pleasure*, *dared not*, or *not like*).

- **Polarity words and expressions.** Words or expressions of positive or negative polarity can help to set the scene with a particular affective mode, in particular with relation to context and acquired knowledge, and expressions of opposing polarity may be used as a contrastive means. Modifiers can additionally intensify the affective load. Note that lexical words and phrases may have *permanent* vs. *occasional* attitudinal meaning and "a word or expression with emotive [here in the sense of polarity or opinionated] meaning can color an entire clause or sentence, so that the clause or sentence expresses approval or disapproval." (p. 151) (Hedquist 1978).

  Sentence examples:

  1. ANGRY-DISGUSTED: *All this was **new** and **interesting**; that is for the first time, but afterwards, as the weather-cock found out, they **repeated** themselves and always told the same stories, and that's very **tedious**, and there was no one with whom one could associate, for one and all were **stale** and **small-minded**.* (WEATHER COCK$_{1,2}$)
  2. FEARFUL: *He will **strike** you **down** before you can cry for **mercy**."* (CORN$_{1,2}$)
  3. HAPPY: *It became a **splendid** flower-garden to the **sick** boy, and his little **treasure** upon earth.* (SICK BOY$_{1,2}$)

[24] This lexeme could arguably be called *affect* word, rather than *affect related*.

[25] Similarly, Mihalcea and Liu (2006) found "love" grams often occurring with *sadness*.

4. SAD: *She felt **dreadfully cold,** for her clothes were **torn**, and she was herself so **frail** and **delicate**, that **poor** little Tiny was nearly **frozen to death.*** (HEROINE$_{1,2}$)

5. SURPRISED: *Who would **trouble** themselves about such **trifles**? especially at a **comedy**, where every one is expected to be amused.* (MAIN CHARACTER$_{1,2}$)

- **Acquired knowledge and human experience.** Knowledge about situations, visualizations, and behaviors which readers take for granted from experience may be associated with particular affects. For example, it is common knowledge that death is traumatic for remaining loved ones. In addition, story worlds involve canonical representations of characters, actions, functions, situations and objects, and surrounding **context** can also be an important source for affective interpretations as shown in the below examples.

  Examples:

  1. **Inspiration from weather, flowers, nature, and God**

     HAPPY: ***Whenever the sun shone, we felt his warm rays,*** *and* ***the little birds would relate stories to us as they sung.*** (MATCHES$_{1,2}$)

     HAPPY: *With **pious gratitude** the girl looked upon this **glorious work of God**, and bent down over one of the branches, that she might examine **the flower and inhale the sweet perfume.*** (GIRL$_{1,2}$)

  2. **Singing (or dancing, jumping)**

     HAPPY: *Here I come!" shouted Jack* the Dullard, and *he **sang till his voice echoed** far and wide.* (JACK$_{1,2}$)

  3. **Growth**

     HAPPY: Then *it seemed as if **new life** was thrilling through every fiber of root and stem and leaf, rising even to the highest branches.* (MAIN CHARACTER$_{1,2}$)

  4. **Physical lack and need**

     SAD: *He was **hungry and thirsty**, yet no one gave him anything; and when it became dark, and they were about to close the gardens, the porter **turned him out**.* (HERO$_{1,2}$)

  5. **Sleep deprivation or allowance**

     ANGRY-DISGUSTED: *"What a buzzing and rumbling there is in the elfin hill," said one of the lizards; "I **have not been able to close my eyes for two nights** on account of the noise; I might just as well have had the toothache, for that always **keeps me awake**."* (LIZARDS$_{1,2}$)

     HAPPY: *A little homely bed was prepared for him, but to him who had so often slept on the* *hard stones it was a royal couch, and* ***he slept sweetly and dreamed of*** *the splendid pictures and of the Metal Pig.* (HERO$_{1,2}$)

6. **Alcohol addiction** (here contrasting with music and scholarship)

   SAD: *They say that he was once an energetic young man, that he studied the dead languages, and sang and even composed many songs; then something had happened to him, and in consequence of this* ***he gave himself up to drink****, body and mind.* (MAIN CHARACTER$_{1,2}$)

7. **Incapability**

   SAD: *"And I, poor fellow," said the drover, "I* ***who am so old already, cannot get there****."* (OLD MAN$_{1,2}$)

8. **Unexpected observation**

   SURPRISED: *The duckling had* ***never seen any like them before****.* (MAIN CHARACTER/HERO$_{1,2}$)

9. **Appearance, posture (or facial expression, intonation)**

   SAD: *The old grandfather* ***wiped his eyes****, for he had known King Frederick, with his silvery locks and his honest blue eyes, and had lived for him, and* ***he folded his hands and remained for some time silent****.* (GRANDFATHER$_{1,2}$)

10. **Contextual guidance**

    SAD: *Song doth not carry* ***them*** *forth over the lands, nor into the hearts of men; therefore I have no rest and no peace."* (SPIRIT$_{1,2}$) [The spirit's deeds are forgotten, and for that reason he cannot find peace.]

    SURPRISED: *As if* ***these things*** *were of any consequence!* (MAIN CHARACTER$_{1,2}$)

11. **Marriage-related**

    HAPPY: *"I* ***will take her****," said the butterfly; and* ***he made her an offer.*** (MAIN CHARACTER$_{1,2}$)

In fact, one could argue that most of the characteristics mentioned in this overall discussion can be traced back to developed knowledge, experience, associations, and context, and this is a great part of what makes the problem of automatic textual affect prediction so challenging; it involves levels of deep cognitive understanding rather than extractable surface features. I address this issue further in Ch. 6 on future work.

- **Speech acts.** Speech acts reflect a particular kind of *communicative knowledge* which can have affective meaning as discussed in Sec. 2.2. The below examples have the same high agreement affective label, but appear to represent different communicative functions.

Examples:

1. **Cursing**

   ANGRY-DISGUSTED: ***Let her be expelled from*** *the congregation and the Church.* (VILLAIN$_{1,2}$)

2. **Insulting**

   ANGRY-DISGUSTED: ***"It is stupid nonsense*** *to allow yourself to shoot out this way; we are not here to support you."* (OTHER PLANTS$_{1,2}$)

3. **Commanding**

   ANGRY-DISGUSTED: ***"Feet off*** *the table-cloth!" said the old goblin.* (OLD GOBLIN$_{1,2}$)

4. **Threatening**

   ANGRY-DISGUSTED: *"Thou child of perdition,* ***I will yet carry out my purpose!"*** *cried the Bishop of Borglum.* (VILLAIN$_{1,2}$)

5. **Accusing**

   ANGRY-DISGUSTED: *You need not be in a hurry;* ***you are always so impatient,*** *and the youngster is getting just the same.* (FATHER SNAIL$_{1,2}$)

6. **Blaming**

   ANGRY-DISGUSTED: ***"You'll kill me*** *with your crowing," she cried,* ***"it's all your fault.*** (HEROINE$_{1,2}$)

- **Types of direct speech.** Direct utterances may be used by characters to express affect, inferred by first or second person tokens,[26] quotes, or imperatives, and possibly introduced by words of speaking (e.g. *ask, cry* (=shout), *exclaim, reply, retort, sang, say, sigh, shout, speak, utter*).[27] For text-to-speech synthesis, given the default expressive style of storytelling (see Sec. 2.5), direct speech marked by affect is likely to have particularly sharp expression of the affect. The selected examples below demonstrate that direct speech involves a bundle of phenomena in this data subset.

Examples:

1. **Direct speech**

   FEARFUL: ***"Bend your head as we do,"*** ***cried*** *the ears of corn;* ***"the angel of the storm is coming; his wings spread from the sky above to the earth beneath.*** (CORN$_{1,2}$)

2. **Speaking excitedly**

   ANGRY-DISGUSTED: *He* ***spoke very excitedly,*** *saying that their evil propensities would not be*

[26] Precht (2000) claimed that *explicit stance*, related to use of first or second person, are more common with modals, stative affect adjectives, and affect verbs compared to other stance categories.

[27] A philosophical question is whether quotes introduced by mental verbs like *think* and negated clauses accompanied by quotation, such as *did not say*, reflect direct speech or not.

destroyed, nor would the fire be extinguished, and they should never find rest. (MAIN CHARACTER$_{1,2}$)

3. **Short utterance**

   SURPRISED: ***“A difference!”*** *cried the sunbeam, as he kissed the blooming apple-branch, and then kissed the yellow dandelion out in the fields.* (SUNBEAM$_{1,2}$)

4. **Interjection, non-lexical speech sounds, or sound effects (single or repeated)**

   Examples: *ah, alas, horray, hurrah, fie, flop, gee-up, good gracious, good heaven(s), goodness me, heaven preserve us, help (me), ho, hallelujah, hurrah, lo, mercy, o God, oh/o-h, sorry, stop, thump, ugh*

   FEARFUL: *“**Mercy!**” cried Karen.* (HEROINE$_{1,2}$)

5. **(WH)-exclamation**

   HAPPY: *“**How beautiful** this world is!” said the caterpillar.* (CATERPILLAR$_{1,2}$)

6. **(WH)-question**

   ANGRY-DISGUSTED: *“Well, then,* ***why do you lie in my way?”*** *she retorted, “you must not be so touchy.* (HEROINE$_{1,2}$)

- **Mixed emotions.** Affective high agreement sentences also include cases of mixed emotions, e.g. affect or affect-related words referring to more than one affect. The “winning” affect may be determined from contextual knowledge and inference, and the contrast might even make the “winning” affect appear more prominent. At any rate, here I remind the reader that this section’s analysis only considered the obligatory affect labels primary emotion and mood, and not any potential optional annotator observations of less salient co-occurring “secondary emotions”.

  Examples:

  1. HAPPY (mixed FEARFUL): *The little boy said not a word; he was half* ***pleased*** *and half* ***afraid.*** (HERO$_{1,2}$)

  2. SAD (mixed HAPPY): *The grand, majestic oak* ***could not be quite happy in the midst of his enjoyment****, while all the rest, both great and small, were not with him.* (MAIN CHARACTER$_{1,2}$)

  3. HAPPY (mixed SURPRISED): ***For a moment*** *the little ones* ***stood silent with astonishment****, and then they* ***shouted for joy,*** *till the room rang, and they* ***danced merrily*** *round the tree, while one present after another was taken from it.* (CHILDREN$_{1,2}$)

  4. HAPPY (mixed SAD): *He now* ***felt glad*** *at having* ***suffered sorrow and trouble****, because it enabled him to* ***enjoy*** *so much better all the* ***pleasure and happiness*** *around him; for the great*

swans swam round the new-comer, and stroked his neck with their beaks, as a welcome. (MAIN CHARACTER/HERO$_{1,2}$)

5. SURPRISED (mixed FEARFUL): *He started back, quite **bewildered** with the **fright** which the goloshes of Fortune had caused him.* (VOLUNTEER$_1$, BIG HEAD$_2$)

6. SAD (mixed ANGRY-DISGUSTED): ***Sorrow and suffering** deprived Anthony's father of his strength, so that he had something else to think of besides nursing his **love-sorrows** and his **anger** against Molly.* (FATHER$_{1,2}$, HERO$_2$)

7. ANGRY-DISGUSTED (mixed HAPPY): *He **did not like** the student, and would **grumble** when he saw him cutting out **droll** or **amusing** pictures.* (LAWYER$_{1,2}$)

8. HAPPY (mixed FEARFUL): *But she only said this to **frighten** Jack the Dullard; and the clerks gave a **great crow of delight**, and each one spurted a blot out of his pen on to the floor.* (CLERKS$_{1,2}$)

Above, I gave examples of some characteristics which seem to appear in the high agreement H.C. Andersen subset. This was a qualitative account and by no means an exhaustive analysis (for instance I did not discuss the use of punctuation marks), but it illustrated the complexity of affect cues in affective sentences marked by high agreement.

Moreover, there may be trends for particular characteristics associating more or less with a particular affect. For example, in this subset, FEARFUL sentences seem to often contain affect or affect related words, whereas SURPRISED sentences may be quite often characterized by various types of direct speech, such as short utterances, interjections, (WH)-exclamations/questions indicating astonishment, or involve unexpected observations. Naturally, these are just tendencies; affect or affect related words do occur with SURPRISED sentences, and do not occur with FEARFUL sentences. However, the following instances illustrate the above points.

1. FEARFUL: *How they **terrified** the poor duckling!* (MAIN CHARACTER/HERO$_{1,2}$)

2. FEARFUL: *"But we are very much **frightened**," said the young storks, and they drew back their heads into the nests.* (MAIN CHARACTER$_{1,2}$)

3. FEARFUL: *Tiny **trembled** very much; she was quite **frightened**, for the bird was large, a great deal larger than herself,- she was only an inch high.* (HEROINE$_{1,2}$)

4. FEARFUL: *So they went on, but the road lead them out of the way; no house could be seen, it grew dark, and the children were **afraid**.* (HERO$_{1,2}$, HEROINE$_{1,2}$)

5. FEARFUL: *He declared, when he quite recovered himself, that this had been the most **dreadful** night he had ever passed; not for a hundred pounds would he go through such **feelings** again.* (WATCHMAN$_{1,2}$)

6. SURPRISED: ***“Understand!*** (HERO$_{1,2}$)

7. SURPRISED: *“**Oh**, indeed; and you will be like her some day,” said he; and he flew away directly, for he felt quite **shocked**.* (MAIN CHARACTER$_{1,2}$)

8. SURPRISED: *“**Good heavens!**” she **exclaimed.*** (WIDOW$_{1,2}$)

9. SURPRISED: *“**It’s a drop of puddle water!**” said Kribble-Krabble.* (MAIN CHARACTER$_{1,2}$)

10. SURPRISED: ***I declare you have put on your Sunday clothes!”*** (JACK$_{1,2}$)

11. SURPRISED: *“**Ho! ho!**” said the old goblin, “**is that what she means?*** (OLD GOBLIN$_{1,2}$)

12. SURPRISED: *“**Heaven preserve us!**” he thought, and **opened his eyes wide,** “**I cannot see anything at all,**” but he did not say so.* (OLD MINISTER$_{1,2}$)

This discussion showed that several interesting characteristics appear to characterize the high agreement affective subset of sentences for the H.C. Andersen subcorpus. It also tentatively hypothesized that some characteristics might show particular affinity with certain affects. Future continued exploration will continue to add to the above insights. Ch. 4 also includes classification results on the high agreement subcorpora.

3.3 Concluding summary

This chapter introduced a new text corpus annotated for affect at the sentence level, and covered labeling schemes, annotation procedures, and subsequent hierarchical treatments and mergers. An interesting subset of high agreement affective sentences at the merged *all* level was discussed in more detail.

In addition, a number of weaknesses can be noted with the corpus. First, this was not a complete treatment of texts by these authors. Second, impact of annotator vs. corpus, or effects of e.g. changing annotator patterns over time was not explored, due to uneven annotation load across several annotators. Variations in annotator behavior and perception need more thorough future comparison. Third, whereas the restricted set of affects at the basic level appeared more tractable for a classification task, the forced-choice label set may not have captured certain facets of affect, and annotator strategies for such cases remains an interesting area for future exploration. Moreover, other textual genres or domains may display different behaviors or require different affect labels. Fourth, there is no guarantee that the annotated affect labels generalize to other readers. In fact, I rather expect other readers to disagree with some of the annotators' assignment of affect labels, in particular for subtle affective cases. Fifth, influence on annotators from also marking other secondary aspects of emotional contents was not explored. Sixth, it is possible that some artifacts were caused by the annotation scheme, procedure, or tool. Future work should add increased control of variables in the data collection setup to facilitate statistical experimentation (e.g. factors influencing annotation), and draw on well-known annotation tools to ensure technical robustness. Seventh, merging mood and primary emotion instead of tie-breaking may also have lead to some disadvantages, e.g. causing sentences with labels of opposing polarity, or some stories may be marked by one dominant affect,[28] for example if the annotator had a low "threshold" for annotating affect and especially when employing NEUTRAL undersampling.[29]

To conclude, this corpus represents a significant first attempt toward creating a resource for affect prediction in text, as well as addressing the intriguing problem of affect perception in text, and more particularly in tales. Several lessons were learned and new research questions emerged for the future. At any rate, this data set is unique in the field. Moreover, the subset of high agreement sentences may be particularly interesting for scholars conducting affect research, whereas other parts of the corpus may reveal insights on affective variation in text and perception thereof.

[28] For example, in *The Old Man and His Grandson* SAD dominates after merging primary emotion and mood for both annotators. However, a few sentences also have other affect labels.

[29] An alternative could be a threshold for adding a label to a sentence; however, it is just as unclear what that would be.

Chapter 4

Affect in text

4.1 Initial overview

The main goal of this chapter is to explore the use of multiclass classification for the problem of predicting affect in text at the sentence level. The rest of this chapter uses the term *classification* in the broad sense of a method assigning one or more labels from a set of possible target labels to a sentence, rather than in the narrow sense of referring to machine learning classifiers, although two of the four main methods presented are based on a machine learning paradigm.

The first section of this chapter introduces the problem formally. The second section deals with classifying sentences in tales at different levels of the affect hierarchy. Next, the methods applied are presented. Whereas the main focus of the methods used in this chapter is on supervised machine learning, with data described in Ch. 3 as labeled examples for training and testing, two additional methods are also considered.

Subsequently, results are presented both for the methods individually as well as after being combined into a joint prediction, using a straightforward combination scheme with three different weighting procedures.

The following section goes a step further by considering more specifically classification given the subsets of sentences for the whole corpus which were marked by high agreement at the merged *all* level, as defined in Ch. 3, as well as NEUTRAL random sampling.

The last section considers the importance of history knowledge for the prediction of a machine learning method, by comparing the use of excluding history of previous sentences' labels in the feature set against including such previous history as features via either previous sentences' true labels or instead considering actual label predictions for previous sentences when testing.

The concluding part of the chapter summarizes the findings and provides a brief discussion for future directions.

4.2 Affect prediction as a classification problem

Determining the emotion of a linguistic unit can be regarded as a multi-class classification problem,[1] with nominal labels as possible classification targets. The problem is expressed formally in Equation 4.1:[2]

$$\begin{aligned}
&\text{Classification step}:\\
&\qquad\qquad T : \text{text}\\
&\qquad\qquad s \in T : \text{sentence}\\
&\qquad\qquad M : \text{set of classification methods}\\
&\qquad\qquad A^i : \text{set of affect labels at a specific level } i \text{ in the affect hierarchy}\\
&\qquad\qquad P : \text{set of predictions for sentence } s\text{, given method } j\\
&\qquad\qquad f(s,j) : s \rightarrow A^i \text{ for classification method } j \in M\\
&\qquad\qquad f(s,j) \subseteq A^i\\
&\qquad\qquad |f(s,j)| \geqslant 1\\
&\qquad\qquad \text{for } j \in M\\
&\qquad\qquad\qquad P(s,j) = f(s,j) \quad \forall s \in T
\end{aligned} \tag{4.1}$$

In words, if s is a sentence in a text T, let A^i be the number of emotion classes $A = \{a_1,\, a_2, .., a_k\}$ at level i, where one member denotes the special case of *neutrality*, or absence of emotion. For each method j in a set M of classification methods, the goal is to determine a mapping function $f(s,j) : s \rightarrow A^i$, such that, for each method, we obtain at least one affect label for each sentence s as prediction $P(s,j)$. For the machine learning methods, a set of features F guide the inference.

As mentioned in Ch. 3, the classification experiments considered three affect levels, see Table 4.1.

Level name (i)	#Labels	Label names (A^i)
all	6	ANGRY-DISGUSTED, FEARFUL, HAPPY, NEUTRAL, SAD, SURPRISED
mid	3	NEGATIVE, POSITIVE, NEUTRAL
top	2	EMOTIONAL, NEUTRAL

Table 4.1: Set of labels at each level of classification, corresponding to the affect hierarchy in Ch. 3

Most methods presented in the next section output one winner label as prediction per sentence. The exception was the *lextag* method, which output any tied winning labels as prediction.

Moreover, in a combination step, see Equation 4.2, different methods' predictions for the sentence s can

[1] At the top level, it becomes binary classification.

[2] Thanks to Rubén Proaño for input on the formulation.

be combined in a single joint prediction, based on weighting the predicted label(s) by each method in one of three schemes, as shown below. The label with the largest sum of winner predictions across methods was selected as overall combined prediction. As for the *lextag* method, the combination step allowed multiple winners (i.e. tied labels).

Combination step:

$$\begin{aligned}
&w_j : \text{weight of method } j \\
&C : \text{set of counters corresponding to } A^i \\
&c_a \in C : \text{counter of affect label } a \in A^i \\
&c_a = c_a + w_j \qquad \forall p \in P(s,j) \text{ and } \forall j \in M \\
&c_{max} = \{c_k \in C : c_k \geqslant c_a, \qquad \forall a \in A^i\} \\
&|c_{max}| \geqslant 1
\end{aligned} \tag{4.2}$$

Combining methods into a joint prediction was motivated by the idea that different methods may support different strengths when applied to the problem. Three straightforward weighting approaches were used. If a method could output more than one prediction, which was the case for the *lextag* method, each prediction of this method was given the same weight, according to the approaches below.

Method name	Weight
liuetaltag	0.1
lextag	0.2
HAtag	0.3
snowtag	0.4

Table 4.2: Weight by method for the *weighted majority vote* combination approach

1. *Simple majority vote:* All methods' predictions were treated equal, so that any prediction was given a score of 1.

2. *Weighted majority vote:* Each method was given a weight, see Table 4.2. Although ad hoc weights were used, a method's weight was chosen to roughly represent the method's importance, based on the performance of the individual methods (corresponding more to the behavior for the *all* level).

3. *Penalty majority vote:* All methods were treated equal except that a NEUTRAL prediction was only given 'half-credit' (i.e. a NEUTRAL weight of 0.5 instead of 1).

4.3 Preliminary insights from early pilot

A couple of useful insights were gathered in early pilot work with a small subset of data labeled with primary emotions, which had been tie-broken in the case of conflicting annotation labels by the author. On the one hand, the pilot indicated that supervised machine learning seemed like a promising approach although features' impact and interaction remained obscure, and, on the other hand, the sequencing of affect across stories and in adjacent sentences seemed useful as knowledge. The pilot work on a subset of 22 tie-broken Grimms' stories showed above baseline performance in binary classification of *emotional* versus *neutral* sentences, with less increase over the baseline in 10-fold cross-validation after applying a conservative versus a liberal tuning/evalution method to determine SNoW's learning parameters. However, since stories may have different overall directions or affective personalities and the affect which occur in texts may differ between stories (e.g. a problem similar to that of having topical subgenres), the effects of randomly selecting tuning data should be systematically explored, as was done in the extended work discussed below.[3]

In addition, hypothesis testing on the tie-broken pilot subset, which considered the ratios of emotion labels in immediately adjacent sentences, showed that for this tie-broken small data set, stories tended to begin NEUTRAL and end HAPPY; that most emotions tended to decay into NEUTRAL; and that after ignoring NEUTRAL adjacent labels and story boundaries different emotions appeared more or less prolonged into adjacent sentences.[4] Moreover, a descriptive examination of ratios seemed to indicate that different polarities might be more prominent in certain portions of stories in the tie-broken data subset. It has been noted before that subjective inclination may differ according to location in the text. For example, Taboada and Grieve (2004) hypothesized that opinionated adjectives tended to occur at reviews' ends. At any rate, it seemed interesting to explore the importance of including history information for labels of previous sentences into the feature set and examine its impact for machine learning methods with more and different data.

4.4 Extending the classification methods

In the continued work, a number of assumptions and methodological decisions were motivated by the affect prediction task itself or the intended application area of affect prediction within the context of children's stories, as outlined below.

- Primary emotion and mood labels assigned by annotators were combined into a set of affect labels (cf.

[3] The pilot classification was run on Red Hat Linux and feature extraction etc. used an earlier python version (probably 2.2). It involved some minor implementation mistakes, and a partial code rerun (e.g. with feature extraction on python 2.4.1) resulted in slightly different results.

[4] For example, SAD and ANGRY appeared less transient, but not FEARFUL and HAPPY, and NEGATIVE emotions showed evidence for the polarity extending into adjacent sentences and POS. SURPRISED into the following sentence.

Ch. 3 for more details regarding the motivation behind this decision). The merged label set with six types of labels were used at the *all* level. Other levels were inferred from original basic labels.

- NEUTRAL labels were *undersampled*, i.e. removed from mixed neutral-affective sentences, given the assumption that affective predictions were preferred in the expressive story domain, to avoid over-training for NEUTRAL predictions.

- Given that emotion perception is subjective and that the concept of "ground truth" does not really apply, predicting any one of a sentence's affect labels was assumed *acceptable* and counted as a success in testing.

- Separating experiments by each of the three different subcorpora and each of the three levels in the affect hiearchy may give an idea about generalization tendencies.

- Stories were given equal weight when computing performance, independent of their length. This meant that computation of mean performance accuracy took place at the story level.

- The exception to the latter point was the high agreement experiment discussed at the end of this chapter. It used leave-one-out cross-validation at the sentence-level, i.e. considering each included sentence individually as a test case in cross-validation experiments and when computing mean accuracy.

Moreover, several different methods were explored, and these are described in the following subsections.

4.4.1 General SNoW learning method: *snowtag*

As in the pilot, the *snowtag* method used a variation of the Winnow update rule implemented in the SNoW learning architecture by UIUC's Cognitive Computation group; for details, see Carlson, Cumby, Rizzolo, Rosen, and Roth (1999), which learns a linear classifier in feature space. SNoW has been successful in several NLP applications, e.g. for semantic role labeling (Koomen, Punyakanok, Roth, and Yih 2005).

Below, I discuss the *attributes* or *feature types* used in the final experimentation for the SNoW methods. As a side note, some of these feature types have a fairly modest set of possible values or features that can be active in a sentence (for example, the feature type *first-sentence* has only one possible active feature or value, i.e. when the sentence is the first sentence in a story the feature would be true and thus active, and it would be false and non-active in any non-story-initial sentence). One could argue that such feature types are advantageous for transparency, inspection, customization, and further experimentation.[5] Some of the other feature types, in particular the *content-word-like sequences*, have a wide range of possible features or

[5] I am grateful for a course with Chilin Shih for classroom discussion on this matter.

values that could be active (e.g. in this case any content-word-like vocabulary sequences of certain length occurring in the text input), vastly increasing the size of the feature set active in the training data. The latter might be more obscure, however perhaps serve for generalization when transporting the feature set to other domains or similar tasks. I leave this discussion open since it extends beyond the scope of a linguistic treatment, but in this context it may be valuable to mention that one of SNoW's advantages is that despite incorporating for example *content-word-like sequences* as a type of feature, average feature vectors in the SNoW input files are rather short because by default they include only the active features for each example instance (compared to some other learners which require a complete feature matrix).

The experiments were run for each of the three subcorpora at each of the three different levels of the affect hierarchy, see Table 4.1 and Ch. 3 for details. The customized experimental setup outlined in Fig. 4.1 considered that the data for tuning may influence outcome, and thus the experiment was replicated in runs with randomly selected tuning data; for results reported below there were 2 runs at each level. For each run, there was a *tuning step*, used to tune SNoW's learning parameters, followed by an *evaluation step* which performed k-fold cross-validation with the parameters resulting from the tuning step. The tuning step split the data of a subcorpus (e.g. B. Potter tales) into two sets with approximately 10% for testing data (8 stories out of 80 for Grimms', 7 out of 77 for H.C. Andersen, and 2 out of 19 for B. Potter) with the rest used for training data. More than 500 different SNoW parameters combinations were considered in the tuning step for each level,[6] and the first of any subset of parameter combinations with the best test result in tuning was chosen for training the SNoW classifier in the evaluation step. Next, the evaluation step included k-fold cross-validation, with k equal to the number of stories in the subcorpus. In other words, $k-1$ stories were used for training whereas 1 story was used for testing in each fold, and this process was repeated for k different test folds. Thus, there was a difference across subcorpora in the number of test folds of the experiments (80 for Grimms', 77 for H.C. Andersen, and 19 for B. Potter), and there was also a

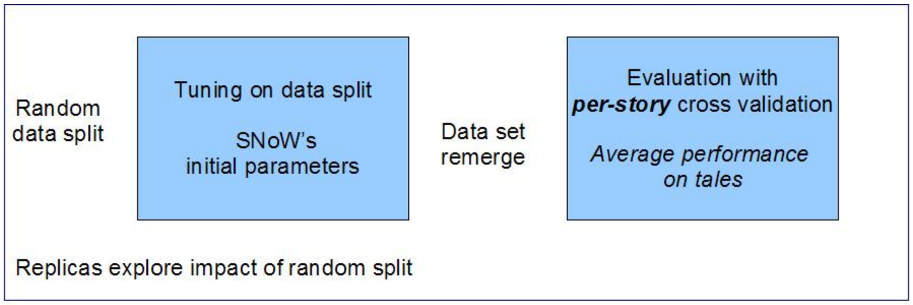

Random data split
Tuning on data split
SNoW's initial parameters
Data set remerge
Evaluation with *per-story* cross validation
Average performance on tales
Replicas explore impact of random split

Figure 4.1: Experimental setup of *snowtag* experiments

[6] SNoW has many paramters, of which only a few were explored.

difference in the number of instances in separate test (and train) folds, since stories were of different length. Although tuning drew on the same data as the evaluation step did, which could have resulted to some degree in overfitting, the data splits for training data and testing data did not correspond in the two steps. Moreover, the tuning phase did not cover feature set experimentation because, combinatorially, this would have been too time-intensive. Thus, it should be acknowledged that there is a risk for having overfitted the feature set on the story data, since throughout the development of the project, some degree of adjustment with features was done on the whole 2-step process (although more so on Potter).[7]

The features for sentence instances were automatically extracted at the sentence level, however true previous labels were drawn from the annotations. The feature types considered were inspired by previous literature to some degree (see Ch. 2) or by the author's familiarization with oral and written stories and linguistics.[8] A feature-id in a sentence or example instance vector in the SNoW input files could represent presence of a feature for the sentence (e.g. the presence of a particular special punctuation mark in the sentence or the sentence being the first sentence in the story) or a discretized representation of a present continuous numerical feature for the sentence (e.g. predefined ranges for sentence length, or by considering the range from the nearest floor and ceiling integers for a lexical score, or counts, ratios, dominance or equivalence). Gupta's python implementation of the Porter stemmer was used for some features.[9] Also, "content-word-like" in the list of feature types below referred to when the first letter of the part-of-speech tag of a token was in the set {N, V, J, R}, i.e. roughly forms of nouns, verb, adjectives, adverbs or particles. Part-of-speech (POS) tags were extracted from trees by Charniak's parser,[10] mostly corresponding to the Penn Tree Bank part-of-speech tag set.[11] Features not active in training were discarded when testing.

The feature types considered in the final experiments for *snowtag* included:

- Poetic: *and-repetitions* ("*w* and *w*" structures); *direct speech* determined by presence of a small set of words (a few words of speaking in the past tense and some 1^{st} and 2^{nd} person pronouns); counts and presence of *double quote, emphasis, onomatopoeia, repetition.*

- Story-related: *% progress* in story, *first sentence* in story, *last sentence* in story

[7] An early development phase inconclusively considered exploring manually grouped feature types combinatorially in tuning. However, such an approach also seemed disfavored because it suffered from the arbitrariness of manual feature type grouping for a still poorly understood linguistic problem. In later experimentation, some features previously explored were excluded (e.g. some of which involved less-accessible external resources, like Whissell's Dictionary of Affect). This appeared to have less influence compared to, e.g. random data split in tuning.

[8] Most features were extracted after some sentence processing. Some noise was expected to have remained or been included in features due to, e.g. sentence processing, feature extraction, or text-peculiarities. For example, poetic feature tagging of sentences during preprocessing drew on simple and incomplete heuristics and were not noise free. Also, ambiguities may not be captured, e.g. quotes might also occur with unspoken thoughts and be used for other functions, and H.C. Andersen stories started on non-emphatic upper-cased words.

[9] See `http://www.tartarus.org/~martin/PorterStemmer`. Some stems may not be adequate, and English singular genitives were not stripped before stemming.

[10] Thanks to Vasin Puyanakok who provided the parses.

[11] Ignoring separate auxiliary tags for verbs like *do, have, be.*

- Orthographic: presence of *special punctuation* (*!*, *?*, stars, three dots)

- Syntactic: *content-word-like POS ratios*, *dominance or equality between pairs of active content-word-like POS tags*, *sentence length*, *UH (interjection-like) tag count*, *presence of word-UH tag pair(s)*

- Content-word-like sequences: presence of *content-word-like sequences* of length one, two, and three.

- Lexical: *presence* and *counts* of words (or stems) from *special word lists*[12] and *dominance or equality between pairs of active special word lists*, with vocabulary that was *affective*, *negative*, *positive*, *interjections*, or *emotion-specific* (e.g. a FEARFUL word list, given certain POS and extension via lexical relations), as well as the combined list of the latter.

- Dimensional lexical scores:

  - Given SentiWordNet's[13] PosScore/NegScore/ObjScore[14] for content-word-like words or stems, *dominance of thresholded polarity* as well as *polarity score sums.*

  - Given word or content-word-like word hits in the ANEW[15] word list (section for all subjects) and their mean ratings for valence or (un)happiness and arousal:*average activation and pleasantness*, *combined representation of average activation and pleasantness*, and *tension* computed as (*average arousal - average pleasantness*)[16]

- Previous history: *True labels of 3 previous sentences* in the story when applicable

4.4.2 High agreement SNoW learning method: *HAtag*

The general idea behind the *HAtag* method was to see what effect training on a set of sentences assumed to represent clearer affective cases would have for predicting over a larger set assumed to include more subtle and subjectively-oriented cases of affect. As for other methods, this approach evaluated on the story-level. (In another experiment, discussed later in Sec. 4.6, a leave-one-out cross-validation experiment was conducted.)

The *HAtag* method used the same learning setup as described for *snowtag* as regards having an initial tuning step and types of feature. Two separate runs were completed. However, in the evaluation step, training and testing were conducted in two phases. Moreover, the tuning step and the first training phase of

[12] Special word lists were obtained in different ways, e.g. from publications, web-browsing, WordNet-based semi-automatic retrieval and extension given certain POS, or inspecting texts. The lists may be incomplete or contain noise.

[13] Version 1.0.1 by Esuli and Sebastiani (2006).

[14] Considering sense #1 and the last occurrence for forms encountered more than once in SentiWordNet.

[15] (Bradley and Lang 1999)

[16] My interpretation of Anderson and McMaster (1982)'s brief description of the tension score may not correspond to their implementation, and they used an alternative lexical resource and human word-sense disambiguation.

the evaluation step drew on a subset of the data fulfilling special criteria, whereas the second testing phase of the evaluation step was performed separately and involved testing on the tales of each subcorpus, given the trained outcome from the first evaluation phase.

Given that the *high agreement* data subset was not large, the special data set included the subset of all sentences marked by high agreement from the whole corpus as well as a NEUTRAL random sample to allow for NEUTRAL predictions. The high agreement sentences, as defined in Ch. 3, included all sentences with four identical affective labels at the merged *all* level, cf. Table 4.1. The NEUTRAL sample was selected at random for each level before tuning from the set of all NEUTRAL sentences. The size of the NEUTRAL sample was determined based on the number of affective labels at that level by Equation 4.3, where HA was defined as the set of high agreement sentences in the whole corpus, and A^i, as defined in Equation 4.1 above, was the set of affect labels at a specific level in the affect hierarchy. Thus, the number of added NEUTRAL sentences equaled a hypothesized uniform ratio of high agreement affective labels.[17]

$$\left\lfloor \frac{|HA|}{|A^i| - 1} \right\rfloor \tag{4.3}$$

Hence, more NEUTRAL labels were added at the *mid* level than at the *all* level, and at the *top* level, the set had an equal amount of EMOTIONAL and NEUTRAL sentences.

After tuning as described above for *snowtag* but with the special data set, the resulting SNoW parameters were used for training on the full special data set. The outcome of the training was then subsequently used to predict sentences' labels at the corresponding level for the stories of a particular subcorpus in a separate testing phase. As for other methods, mean accuracy was computed at the story level across stories.

Since this method involved tuning and then training for evaluation on the high agreement sentences (1207 in total) and a subset of the NEUTRAL sentences, and then tested for evaluation on the tales, a part of the testing data had been seen in training. However, it is also important to bear in mind that the tuning step and the training phase of the evaluation step were not subcorpus specific here. Moreover, as noted in Ch. 3, most affective sentences were not high agreement sentences, but had mixed affective-neutral labels in the original annotations.

4.4.3 Vocabulary method: *lextag*

The *lextag* method used the same special word lists for specific emotions introduced above for the feature types for *snowtag* and *HAtag* methods. However, the *lextag* method used a straightforward heuristic based

[17] Given that the floor was used, for a very small $|HA|$, this could hypothetically result in zero added NEUTRAL sentences. This did not occur, given the size of $|HA|$ for each subcorpus.

on these word lists to assign affect labels to sentences. It counted the number of word or alternatively stem hits for extracted content-word-like words in a sentence, and output the label or tied labels with the maximum number of hits for a given sentence. A default NEUTRAL label was output in the case of no hits.

For ANGRY-DISGUSTED, word lists derived separately had been combined into a joint list, as was also done in the *snowtag* and *HAtag* feature set. In the case of overlap across word lists, a word/stem hit could count towards more than one emotion label.

Additionally, for the *mid* and *top* levels, word lists were mapped in logical relation to the affect hierarchy outlined in Ch. 3. For example, hits in the special word lists for ANGRY-DISGUSTED, FEARFUL, SAD, and SURPRISED word lists counted towards the NEGATIVE label at the *mid* level, and hits for HAPPY and again SURPRISED word lists towards the POSITIVE label. Similarly, hits in any of these special word lists would count towards the EMOTIONAL label at the top level.

It is possible that the size of the word lists affected the outcome of this method, since the lists were of different length, e.g. the SURPRISED word list was notably shorter than the other word lists.

4.4.4 Common sense method: *liuetaltag*

Compared to other methods, the *liuetaltag* method drew on an external implementation, namely the `guess_mood` option in ConceptNet, developed by Liu and Selker. Unfortunately, the implementation partially remained obscure because of the developers' inaccessibility. Nevertheless, the method was included despite these difficulties, on the basis that Liu and colleagues' work on affect sensing is strongly cited in the literature, and the goal of incorporating the *liuetaltag* method was to draw on research based on common sense knowledge data, as it seems that such information is important for affect interpretation (cf. Ch. 3).

The method called *liuetaltag* used the predictions provided by `guess_mood` implemented in ConceptNetNLTools.py distributed with ConceptNet v.2.1 (Liu and Singh 2004), which output a score in the range from 0.0 to 1.0 for each of six Ekman emotions (ANGRY, DISGUSTED, FEARFUL, HAPPY, SAD, and SURPRISED) standing for the relative affinity of a particular mood in the input text.

Each sentence was thus individually evaluated and tagged at each affect level, fitting the method into the label and classification framework described above. As for all other methods, the performance was computed against the actual affect labels at that level, which had been adjusted as described above. The method ran once and its running time was notably slow.

Because of the properties of `guess_mood` and its sparse documentation, a few assumptions were made.[18] The first emotion in the `guess_mood` output, having the highest score, was chosen as winner for the sentence,

[18] Some awkward and different output behavior was also noted across system types for this method.

and if this score was below an ad hoc threshold of 0.25, the sentence was tagged as NEUTRAL.[19] Thus, as with *snowtag* and *HAtag*, this method output one predicted label per sentence and level. Also, since the method's output gave separate scores for ANGRY versus DISGUSTED, any one of those emotions as winner triggered the joint ANGRY-DISGUSTED tag. Lastly, since the output did not consider the valence of SURPRISE, the prediction of this emotion was assumed to always represent a positive stance at the *mid* classification level, based on the simplifying assumption that from the point of view of common sense surprise may be prototypically more associated with positiveness.

4.5 Comparison of the performance of individual methods and combined approaches

The performance of each of the levels for each subcorpus was explored individually by considering both the performance of individual methods as well as on the combined joint prediction, given the three combination approaches.[20]

First, a few words may be needed about the practical implementation before proceeding to discuss results. Each of the methods was first run in isolation for a given level, and the combination step took place on the produced output of annotated stories, separated by level and subcorpus. Specifically, it is important to note that there was a difference in how the set of marked-up stories were produced in the isolated runs, depending on whether methods involved training or not, and whether k-fold crossvalidation was used. To be more concrete, since two of the methods *liuetaltag* and *lextag* did not involve training these two methods simply received each story in a subcorpus as input, and then produced it marked up for affect as output. On the other hand, the two machine learning methods *snowtag* and *HAtag* involved training. For *HAtag*, in the evaluation phase after training was completed on a special subset of the data for a particular affect level, the resulting classifier network was then used to predict labels for each story in a subcorpus for that affect level, i.e. the mark-up process took place in a similar fashion to the above two methods, once a network had been obtained. However, for *snowtag*, which involved k-fold cross-validation experiments, each cross-validation fold resulted in a single marked up test story, and a complete set of marked up stories by *snowtag* was thus obtained only after the full k-fold cross-validation experiment had been completed; thus in a sense, the individual stories for *snowtag* were marked up by multiple instantiations of the *snowtag* method (each trained and tested on different splits of the data). The individual methods' results represent the mean

[19] Ties for winners were not considered for *liuetaltag*, as the first emotion in the output was always chosen.

[20] Experiments were run on Linux Suse with python 2.4.1 used for most implementation, and Neelay Shah's pySNoW interface, see `http://l2r.cs.uiuc.edu/~cogcomp/software.php`.

accuracy across stories after a complete set of marked up stories had been obtained for each method, and for *snowtag* and *HAtag* the span of the mean accuracy for the two runs are included below.[21]

Lastly, the combination step was run after the full set of marked up stories had been obtained for each of the methods. For *snowtag* and *HAtag*, the best performing run was chosen at each level to represent those methods in the combination step. As noted above, in the case of tie-breaks for winner in the combination step, all tie-broken labels were output as acceptable winners, and the average of winning labels for a sentence in the combination step is included below in parenthesis.

As noted above, concerning the decision on *accuracy* of predictions: since a sentence might have multiple true labels, a marked up sentence counted as *correct* or *accurate* if any of its predicted labels was in the set of the true labels (after the NEUTRAL undersampling described before). This measurement of accuracy corresponds to the notion of *acceptability* discussed before; given multiple *acceptable* answers, it suffices to get at least one of them right.[22]

In addition, two baseline methods were reported upon in the comparison (also computed after NEUTRAL undersampling, as done for the methods). The first baseline, *N-BL*, was the mean ratio of NEUTRAL sentences across stories, whereas the second baseline, *Freq-BL*, captured the mean ratio of the most frequent label across stories. Whereas the two baselines were the same for Potter, they differed for Grimms, and also differed for H.C. Andersen at the *top* level.[23]

		Baselines		Individual classification methods				Method combinations		
		N-BL	Freq-BL	*liuetaltag*	*lextag*	*HAtag*	*snowtag*	simple	weighted	penalty
Potter	all	51	51	20	51	59-60	58-**60**	65 (1.2)	63 (1)	**66** (1.4)
	mid			31	54	60-63	**64-67**	69 (1.2)	68 (1.1)	**70** (1.2)
	top			48	56	57-63	**71**	**81** (1.3)	79 (1.2)	66 (1)
Grimms	all	24	24	22	36	52-56	**58-59**	62 (1.4)	59 (1.1)	**66** (1.5)
	mid		55	42	40	54-60	**66-67**	69 (1.2)	69 (1.2)	**73** (1.3)
	top		76	68	46	48-57	**78**	**86** (1.4)	84 (1.3)	76 (1)
HCAnd	all	40	40	22	44	55-58	**60-62**	65 (1.2)	62 (1)	**67** (1.4)
	mid			36	47	58-61	**67-68**	69 (1.2)	69 (1.1)	**70** (1.2)
	top		60	57	52	53-60	**69-72**	**83** (1.4)	80 (1.3)	69 (1)

Table 4.3: Mean accuracy across tales for individual methods and method combinations per subcorpus. For the combined predictions, the average number of joint labels per sentence is in parenthesis.

The results in Table 4.3 indicated that although the numerical performance differed across levels and subcorpora, and improvement over the baselines were noted for some but not all methods given a particular subcorpus and level, the performance trends tended to be similar for individual methods across subcorpora. Most methods increased in accuracy when moving up the affect hierarchy, linked to the shrinking size of the

[21] Here, *HAtag* was run twice with run-related randomness for tuning and selection of NEUTRAL sample.

[22] This is SNoW's default: "SNoW will count the example as correct if any of the targets that appear in the example are predicted." (SNoW manual). Specifically, its 'winners' output option was used when tagging sentences.

[23] The sum of all ratios for labels could exceed 1, since a sentence could have more than one label.

confusion set of possible target labels.

Moreover, the best performing results for individual methods as well as for the combination step are marked in bold-faced font and show that for individual methods the *snowtag* method generally performed best, given the feature set which involved the true previous labels, mostly followed by the other machine learning method *HAtag*. The latter method showed more performance range; i.e. it tended to perform similar or worse at either the *mid* or *top* level, compared to lower levels, likely due to the run-specific randomness.

The effect of randomly selecting data was also notable in the span of mean performance for *snowtag* and *HAtag*. Thus, since the *snowtag* and *HAtag* methods involved randomness, the results in Table 4.3 should be regarded as indicative, rather than fixed, especially since during development results may have fallen inside as well as outside the above ranges. The results for other methods should probably also be regarded as indicative, since they depended on external resources (word lists) or specific thresholds. Moreover, since the results reported are average accuracies, it is important to note that the performance variation was high between individual stories; the performance could be good on individual stories and notably poor on others. Most of these variational effect were expected given the subjectivity involved when collecting the data set.[24]

In addition, the combined results indicated that joint prediction based on combining individual methods' predictions seemed useful for this task. However, as one would expect, this also appeared somewhat linked to increasing the average number of labels per sentence; especially for the *top* level with its two possible labels (EMOTIONAL or NEUTRAL). Moreover, the usefulness of combining methods into a joint prediction also seemed to depend substantially on including sufficient methods.

4.6 Exploring the high agreement subset

As noted in Ch. 3, an interesting subset of the overall corpus was 1207 sentences marked by high agreement, meaning that all four affect labels assigned by human annotators represented the same affective label for the sentence, given the affective labels at the *all* level (see Table 4.1). As discussed above, it was hypothesized that this special subset of the data might represent more clear-cut affective data, and that it would be interesting to explore its classification accuracy. To avoid any confusion with the *HAsnowtag* method presented above which evaluated at the tale level, the experiment described in this section is referred to as *LOOHAsnowtag*[25] Note that for this experiment, the experimental methodology was mostly the same as for *snowtag*, with true previous labels included as previous history features in the feature set, but the data and the cross-validation method for the evaluation step differed, as described below.

[24] For another task with stories, the ESPER system also noted high performance variation across two stories for speaker identification (Zhang, Black, and Sproat 2003).

[25] The abbreviation captures the leave-one-out highagree (LOOHA) scenario.

The data consisted of the high agreement sentences in the whole data set as well as a random sample of NEUTRAL sentence example instances added at each level, as shown above in Equation 4.3. Due to the smaller size of the high agreement data subset, a **leave-one-out** cross-validation experiment was conducted. The NEUTRAL sample was selected before the two runs with tuning and evaluation steps. Thus, the impact of random selection of tuning data could also be examined.

The leave-one-out setup at the evaluation step meant that given a data set of n sentences, at each cross-validation fold, $n-1$ sentences were used for training and one sentence for testing, with this process being repeated over n different test folds. Table 4.4 below contrasts the mean cross-validation accuracies of the two baselines and the method *lextag* against that of *LOOHAsnowtag* on this task. The *lextag* method selected a separate NEUTRAL sample in each of the two runs.

		Data size	Baselines		Individual classification methods	
		#Sentences (#N sent. sample)	N-BL	Freq-BL	*lextag*	*LOOHAsnowtag*
All data	all	1448 (241)	17	31 (HAPPY)	54-55	**69-70**
	mid	1810 (603)	33	40 (NEG)	60	**69-73**
	top	2414 (1207)	50	50 (any label)	69	**79**

Table 4.4: Mean accuracy across sentences in high agreement experiment, given the data subsets taken from the whole corpus (span of mean accuracy over 2 runs)

As seen in Table 4.4 the effect of random tuning data selection was again noted, as shown in the performance span of *LOOHAsnowtag* over the two runs. Moreover, the results for *lextag* showed that the random selection of the NEUTRAL sample also affected the outcome.[26] The above results should be regarded as indicative, and not as fixed, given the random elements in the experimental setup.

More importantly, the results in Table 4.4 showed that above baseline performance was achieved for both classification methods, and additionally also that *LOOHAsnowtag* performed better than *lextag*. Moreover, the results on this special data set tended towards slightly better performance compared to individual methods on the subcorpora in the previous section, although such comparison requires caution, given the differences in the data sets and the cross-validation setup. The above results support the intuition that this data represent more clear-cut examples of affect in text.

4.7 Exploring the importance of previous affect history

Lastly, given that the previous history information appeared important, the impact of including or excluding history among the feature set was explored for the *snowtag* method. This experiment compared when the previous history was excluded all-together as feature type against when previous history was included as

[26] During development, somewhat lower results were also noted for *LOOHAsnowtag*, as well as performance ranges for different levels for *lextag*.

feature type. Moreover, for the latter case, two conditions were explored; either the true labels of the three previous sentences or (during testing) the actual predictions for the same amount of previous sentences were included as features. For both cases which included previous history, training used the true previous labels as features. Again, performance ranges across two runs are presented, and should be regarded as indicative due to the randomness embedded in the experimental setup. For the true history condition, the same runs as above in Table 4.3 are included below.

		Baselines		*snowtag* methods		
		N-BL	Freq-BL	true history	predicted history	no history
Potter	all	51	51	58-60	42-50	50-51
	mid			64-67	51-53	58-59
	top			71	59-61	60-62
Grimms	all	24	24	58-59	33-35	36-39
	mid		55	66-67	56-57	57-58
	top		76	78	76	76
HCAnd	all	40	40	60-62	42-43	47
	mid		40	67-68	49-50	48-51
	top		60	69-72	64	62-63

Table 4.5: Mean accuracy across tales for affect history experiments (span of mean accuracy for 2 runs)

The results in Table 4.5 showed that whereas including previous history in the form of true labels was beneficial compared to excluding the history feature, on the contrary a deterioration in performance was observed when instead adding actually predicted previous labels when testing. In fact, *snowtag* often performed worse with predicted previous labels, compared to when excluding history in the feature set.

This finding indicated that excluding the previous history feature type would probably seem a wiser decision, since true labels would not be available for untagged data. However, one should also bear in mind that when excluding previous sentences' labels from the features, the gap to the baselines as well as other untrained methods (cf. Table 4.3) either shrunk or equalized, depending on the level.

Nevertheless, it is important to note that the results for the feature set with true previous labels in testing showed that history information seemed beneficial for this data set. The discrepancy most likely lay in that the actual history predictions produced for testing were, overall, at present not accurate enough to add benefit.[27] However, with overall more accurate predictions in future development of affect prediction in text, one may expect actual predicted previous history to surface as practically more useful for feature set inclusion.

[27] Also, the truth could have more than one label per sentence.

4.8 Concluding summary

To conclude, this chapter discussed the use of classification for tackling the problem of automatic affect prediction in text. After presenting a multiclass and multilevel statement of the problem with an additional combination step as approach toward this classification task, as well as introducing and motivating a number of assumptions, each method was presented separately. Next, performance results for individual methods, as well as for combined predictions after applying various combination schemes, were shown. The results showed that whereas it may be less common practice to do so in reports on NLP tasks, it is important to examine the impact of, for example, the random split of the tuning data, and evidence of random impact was also found when selecting a NEUTRAL sample to accompany the high agreement data. These effects surfaced despite a modest numbers of runs.

Given the current implementation, when comparing individual methods, the SNoW-based machine learning methods often appeared more successful. However, these methods generally drew on true previous history, and one should also consider that if settings and resources used for the other methods would have been customized or empirically optimized, their performance might possibly have improved, e.g. if *lextag* lists had been lexicographically improved, or if the NEUTRAL cutoff threshold for *liuetaltag* would have been set through an optimization process.[28]

Moreover, the leave-one-out experiments involving high agreement and selected NEUTRAL sentences indicated quite encouraging results at each level, with best results for the SNoW-based machine learning method including the true previous history. This corresponds to the intuition that high agreement sentences were more straightforward for humans to assess as having particular affective meanings, and thus, such data might represent clearer cases to train and test on.

In addition, the experiments with previous history indicated that history seemed useful if representing the truth, but otherwise less so at the present state. In general, for the machine learning methods, the impact of features and how they interact remained fairly unexplored topics, deserving more investigation. This aspect and other suggestions for future directions for predicting affect in text are covered in Ch. 6.

Lastly, I return to a point which, although not new, seems interesting for continuing a broader discussion of computational linguistic research involving subjective natural language processing tasks. The task empirically explored in this chapter brings to question the concept of "ground truth" or "gold standard" as arguably flawed. There may be a need for reassessing the validity of such terminology as well as its implications for how these language phenomena are studied and evaluated computationally, as further discussed in Ch. 6.

[28] Naturally, increasing customization may not transport to other data sets.

Chapter 5

Affect in speech

Speakers express a tremendous amount of communicative information not only in utterances' content, but also in the way the message is conveyed. Whereas the former refers to the verbal message, the latter involves *paralinguistic* cues, such as voice inflection, facial expressions, gestures, and so on. This paralinguistic information helps listeners to make sense of the linguistic message within a conversation's social context. For example, paralinguistic voice inflection can guide a listener to interpret the functions of spoken utterances, and create a communicative picture of interlocutors or parties in the conversation.

A crucial dimension of paralinguistics is affect or emotion, and this chapter focuses on affect in speech. More specifically, it empirically explores duration, pitch and intensity in emotional speech. These are fundamental parts of listeners' perception of a large package of voice inflection cues, which is generally jointly referred to by the term *prosody*. As noted in Sec. 2.4, prosody is used not only for expressing emotional information, but also to provide linguistic cues such as lexical word stress, lexical tone, discourse signals, and so on. The multitasking of prosody is part of what makes specific prosodic phenomena, such as emotion in speech, so challenging to investigate. However, computational applications or interfaces with speaking artificial agents cannot be natural conversation partners without adequate prosodic behavior, and implications also extend to areas beyond speech technology, as suggested in Ch. 1.

Ch. 2 showed that research in speech, prosody, and emotion is a broad and vibrant research area, and it also gave an overview of the field, noting that there are many complications of affective speech and prosody, especially in terms of linguistic or algorithmic complexities.

The work in this chapter[1] is based on Genetic Algorithms, a technology which allows an experimenter to examine current problems of emotional prosody as well as begin asking interesting new questions about affective speech. By conceptually visualizing the problem as one of optimization, which can draw on human preference judgements, this approach treats emotional prosody as something that evolves via interactive feedback from listeners. Moreover, GAs may allow investigators to go beyond looking at what are common patterns in speech databases, and ask interesting questions, such as if there are prosodic patterns that are

[1] This chapter partially contains previously presented material from (Alm and Llorà 2006).

preferred by listeners for different emotions, and how prominent preferential *variability* is. In addition, given the flexible encoding of a search method, an experimenter can frame the studies to, for example, explore word-level or even more *local* prosody. This does not mean to discount the importance of global prosody; a focus on local prosody does not mean that it is sufficient for expressing affect, but it extends prosodic exploration into wider realms. At the same time, when conceptualizing emotional prosody as a search problem, one ought to consider that an exhaustive search of listeners' perceptual space may not be feasible. Using a sophisticated search method allows one to address emotional prosody from a quantitative point-of-view within a reasonable time frame. Another advantage of using *interactive* search with human judges is that one can modify prosody within the course of an experiment in flexible ways dependent on the user feedback, taking into account that perception of affect is subjective. Also, given interactive speech modification, e.g. with resynthesized (computer-modified) speech, the experimenter is not limited by access to expensive speech databases, or to the emotions covered by a given speech database.

The rest of this chapter describes the pilot study[2] as well as the extended study conducted for emotional acoustic search. However, before discussing the experimental studies, the next section first briefly covers the general idea of *Genetic Algorithms* (GAs) with the example of the Simple GA, and then a section introduces the *active iGA* (aiGA). Next, I describe the experimental studies, including setup, participant feedback, results, and summarizing conclusions. I start with the pilot study, and then move on to discuss the extended study.

5.1 Background on Genetic Algorithms

GAs are search algorithms that draw on basic concepts from evolution; for overviews see (Goldberg 1989; Goldberg 2002). Before discussing the particular method employed in the following two experimental studies, the following illustrates the main ideas behind this family of metaheuristics with the example of a simple GA (SGA).

An SGA involves an iterative search process that resembles natural adaptation phenomena, where iterations are termed *generations*. It starts with the encoding of a set or *population* of *individuals*. Individuals have *chromosomes*, consisting of parts or *gene alleles* which represent different pieces of the problem, for example as bit strings or real numbers.

At the initial generation, a population of individuals and their gene values are initialized, usually at random. Next, in an iterative search procedure, individuals from the population are evaluated on the basis of an *objective function*, which is a goodness measure that computes the *fitness* or goodness of each

[2] The pilot study was presented at IS2006. Some modifications have been made.

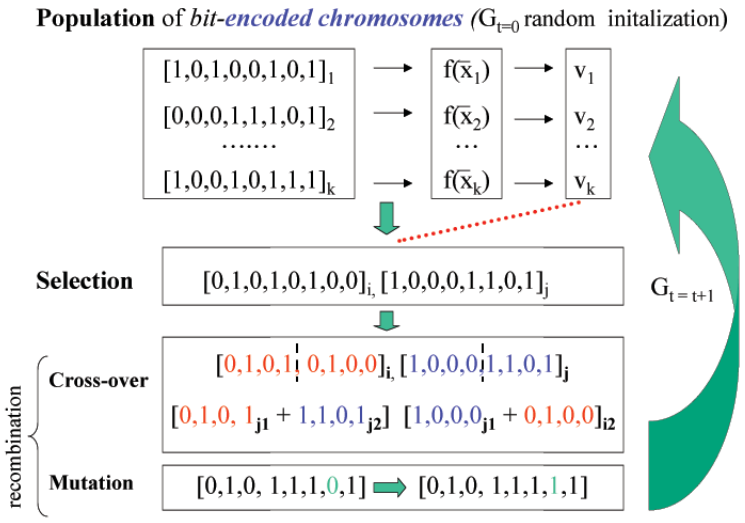

Population of bit-encoded chromosomes (G t=0 random initalization)
[1,0,1,0,0,1,0,1]1
[0,0,0,1,1,1,0,1]2
....
[1,0,0,1,0,1,1,1]k
f(x1)
f(x2)
...
f(xk)
v1
v2
...
vk
Selection
[0,1,0,1,0,1,0,0]i, [1,0,0,0,1,1,0,1]j
G t = t+1
recombination
Cross-over
[0,1,0,1 0,1,0,0]i, [1,0,0,0 1,1,0,1]j
[0,1,0, 1j1 + 1,1,0,1j2] [1,0,0,0j1 + 0,1,0,0]i2
Mutation
[0,1,0, 1,1,1,0,1]
[0,1,0, 1,1,1,1,1]

Figure 5.1: Example of the iterative process of a Simple Genetic Algorithm with bit encoding

individual. This evaluation function could be implemented automatically, e.g. as a minimizing cost function, or interactively, with human subjects performing evaluation of samples. In the latter case, it is important to address issues such as user fatigue and dynamic targets, for example by introducing a *synthetic fitness model* that takes decisions extending beyond the limited input which users can provide since humans get exhausted quickly from such monotonous judgement tasks. The result is that at each generation, parts of the population are selected as parents for the next generation, based on fitness evaluations and generally some form of selection bias, so that more fit individuals are more likely to reproduce. For an SGA, the next generation is usually generated by applying *genetic operators* to the selected parents, which modify the parents in one way or another in order to breed children or new individuals with the hope to increase fitness. The process is repeated over generations until a stop condition is reached. The stop condition could be for example a certain fitness criterion, a specified number of iterations, or something else.

Given a specific application area for GAs, a number of issues need to be addressed, such as GA type, genetic encoding, design of the fitness function, and so on. A characteristic of GAs is that their parameters and characteristics need to be tuned for each particular problem instance. Moreover, certain issues specifically relate to *interactive evolutionary computation* (Takagi 2001) and are relevant for *interactive GAs* (iGAs) which draw on humans' subjective evaluations.

This short overview intended to give a rough idea of what GAs can be like, however it also ended with the fact that there are several consideration which needs to be balanced when using GAs. The next section introduces the particular GA used in the empirical experiment to evolve emotional prosody.

5.2 Evolving emotional prosody by aiGA

Emotion is expressed by prosodic cues, but exploring their interplay and combinations is complicated by a challenging search space. This work evolves emotional prosody in perceptual resynthesis experiments with the *active interactive Genetic Algorithm* (aiGA); for technical details see (Llorà, Alías, Formiga, Sastry, and Goldberg 2005). Within the aiGA framework, fundamental parameters of emotional prosody are evolved by searching the perceptual space of listeners via their interactive feedback based on listening to utterances. These judgements serve to guide the evolution of the aiGA's models, as permitted by imposed bounds in the experimental setup. The aiGA thus evolves a synthetic model beyond what was presented to the user.[3] The aiGA method has been successfully applied to speech before, by interactively estimating cost functions for unit-selection text-to-speech synthesis (Llorà, Alías, Formiga, Sastry, and Goldberg 2005). This algorithm can speed up convergence time, which relates to the size of the problem.[4]

In these experiments, the aiGA evolved variables encoding values for prosodic parameters, used for resynthesizing (i.e. computationally modify) utterances which were presented to listeners. The algorithm assumed variable independence and evolved a probabilistic model, based on a population of normal distributions[5] with the $UMDA_c$ algorithm (Gonzáles, Lozano, and Larrañaga 2002), as implemented by Llorà.[6] The solution of each run r was an evolved synthetic normal model (μ_r, σ_r) for each prosodic variable.

After having introduced GAs as search procedures and the aiGA method for interactive search with human judges, the next section goes on to discuss the details of the experimental pilot study for the problem of evolving emotional prosody for two emotions, namely SAD and ANGRY in one-word utterances.

5.3 Pilot study

This section covers the pilot study, including experimental setup, participant feedback, and results and discussion. The pilot study dealt with evolving emotional prosody for two emotions, SAD and ANGRY, in utterances consisting of one word with the aiGA which was implemented by Llorà. As in the later extended

[3] For aiGA, subjective evaluation based on cognitive targets in participants' mind guide the search process, which models a *synthetic fitness* by taking into consideration partial graph ordering of evaluated individuals. Ordering is obtained by binary tournament selection, where subjects indicate which of two presented solutions is better, or that no solution is better or they are equally bad, and a complete order can then be inferred. The graph guides against contradictions by breaking cycles by maintaining later added edges, and then ϵ-SVM regression estimates the synthetic fitness model, stopping when sigma is below a certain threshold. At each iteration, the population grows. Half of the individuals go on to the next generation of subjective evaluations, whereas half is replaced by new individuals based on the evolved synthetic model.

[4] In the study by (Llorà, Alías, Formiga, Sastry, and Goldberg 2005), aiGA ensured high consistency in subjective evaluations within runs, avoiding contradictive judgements, and decreased user evaluations compared to a simple iGA, i.e. combating user fatigue.

[5] If listeners agree cross-culturally on recognizing emotions above chance for emotions like ANGRY vs. SAD speech, this supports assuming normality. At the same time, a weakness is that this assumption is qualitative and based on perception rather than production. The true distribution might also differ by emotion.

[6] For specifics on the $UMDA_c$ implementation, see Dr. Llorà's home page.

study, the participants listened to and evaluated human speech which was automatically resynthesized (i.e. computer-modified) using the Praat tool. The resynthesis parameters depended on the evolutionary process of the aiGA.

5.3.1 Experimental pilot setup

In the pilot experiment, users listened to and evaluated pairs of resynthesized utterances. The aiGA's population consisted of parameter vectors or *individuals* with information on how to resynthesize a one-word NEUTRAL utterance. Thus, each individual held three gene values for modifying Int (intensity), $F0M$ (mean F0), and Dur (word duration). Other prosodic parameters, e.g. F0 contour shapes or pausing, were not captured in this setup. The search space, see Table 5.1, was adjusted to this case to avoid unnatural speech. As shown in Table 5.1, in the pilot study, *Int* and *F0M* bounds were set to fixed values, whereas *Dur* bounds were encoded as ratios.[7]

Each run had three generations, found sufficient in earlier work (Llorà, Alías, Formiga, Sastry, and Goldberg 2005), and in each generation, a user evaluated a set of pairs of sounds. In total a user evaluated 22 pairs of sounds in a run. Individuals were initialized randomly with a different seed for each day of the experiment, except one individual's values were set according to trends in the literature for each emotion. The fixed individual intended to additionally speed up convergence and help $UMDA_c$, which performs best when starting in closer proximity of the optimum (Gonzáles, Lozano, and Larrañaga 2002).[8]

Variable	Unit	Min	Max
Sound *Int*	dB	69	83
Mean *F0M*	mel (Hz)	124 (139)	227 (282)
Total *Dur*	ratio	0.70 (shorter)	1.80 (longer)

Table 5.1: Upper and lower limits on word-level prosodic properties in pilot

The aiGA encoded prosodic variables in $[0, 1]$ range, where 0.5 represented the NEUTRAL original sound used as resynthesis basis, 0 corresponded to the minimum, and 1 to the maximum allowed. Whether a gene value was greater or smaller than 0.5 related to how that prosodic variable increased or decreased, given the original NEUTRAL word's parameters at 0.5 and the bounds. When resynthesizing an individual, the aiGA's input to the resynthesis was converted to numerical values for resynthesis of prosodic variables (the aiGA truncated input to the resynthesis if the synthetic model evolved beyond $[0, 1]$).[9]

Each user-evaluation was a tournament that involved listening to two resynthesized 1-word utterances,

[7] The pilot's original NEUTRAL words came from utterances recorded in a sound-proof booth in UIUC's phonetics lab, and values ranged between 71-79 dB, 163-204 Hz, and 0.4-0.9 s.

[8] Values for fixed individuals in the pilot study were [0.2,0.2,0.8] for SAD and [0.8,0.8,0.2] for ANGRY for *Int*, *F0M*, and *Dur*.

[9] The expressions below show the procedure used in the pilot for converting between actual values for resynthesis and proportion encoding in $[0, 1]$ range, where N_j is the value of an original NEUTRAL sound j and 0.5 the NEUTRAL proportion, and

and selecting the one which the user felt best portrayed the target emotion, or indicating a draw. To further avoid user fatigue, the word for a tournament was chosen at random from a set of four NEUTRAL words, then resynthesized given two individuals' parameters, and the resulting pair of sounds were presented to the user. The original words used as resynthesis basis in the pilot, see Table 5.2, came from NEUTRAL declarative utterances recorded from a female US English speaker. Words were controlled for length in number of syllables but not for segmental makeup.

Monosyllabic	sas, bem, face, tan
Bisyllabic	barlet, person, cherry, tantan
Trisyllabic	bubelos, strawberry, customer, tantantan

Table 5.2: Words used as resynthesis basis in pilot by syllabic type

Praat is a freely available signal processing tool which is fairly common among experimental phoneticians. Resynthesis was done with two parallel Praat (Boersma and Weenink 2005) implementations,[10] and individuals were resynthesized on the fly in a step-wise modification process (with a final concatenation-resynthesis component) before each tournament with the aiGA running in the background and regulating resynthesis parameters, user tournaments, and computation of evolving performance.[11]

Interactive experiments involved two males and two females; all highly proficient in English, with either Swedish (3) or Spanish (1) as native language. Over ten different days (within a 20-day period), they completed two blocks of three SAD tasks and three ANGRY task, with an intermediate short break, for each day, i.e. SAD and ANGRY target emotions in combination with either *monosyllabic*, or *bisyllabic*, or *trisyllabic* word types.[12] The ten day replicas were done to not overload users and keep them alert, and to reduce effects from random initialization or daily moods. Emotion perception is subjective, so averaged results across runs are of main interest. The tournaments, where users clicked on icons to listen to a pair of resynthesized sound files, were displayed with a customized DISCUS web interface provided by Llorà, with icons displaying emotions and sounds, see Fig. 5.2.

from proportion encoding to actual resynthesis values. I am grateful to Rubén Proaño for deriving these equations.

$$p_j(v_j) = \begin{cases} \frac{v_j - v_{min}}{N_j - v_{min}} \times 0.5 & v_j \leqslant N_j \\ (\frac{v_j - N_j}{v_{max} - N_j} + 1) \times 0.5 & v_j > N_j \end{cases}$$

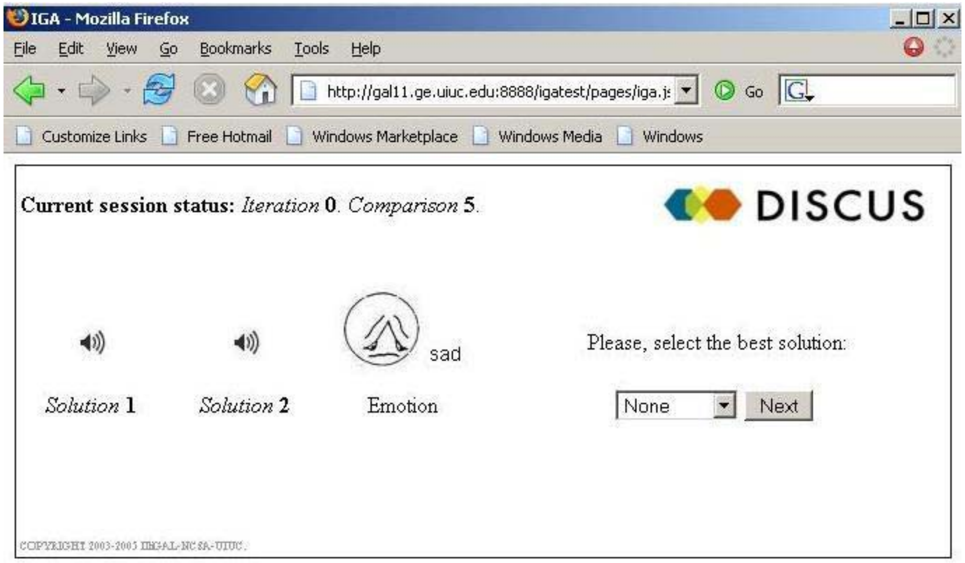
$$v_j(p_j) = \begin{cases} p_j \times \frac{N_j - v_{min}}{0.5} + v_{min} & p_j \leqslant 0.5 \\ (\frac{p_j}{0.5} - 1) \times (v_{max} - N_j) + N_j & p_j > 0.5 \end{cases}$$

Because this procedure takes on different numerical values for different words, it seems to make most sense for *Dur*. Thus, in the extended study, it was only used for *Dur* and the other variables were computed with fixed step sizes.

[10] Only *one* active iGA was running in the background; the parallel resynthesis does *not* mean that a parallel iGA was used.

[11] Praat resynthesis had some *F0M* or *Int* inconsistencies; assumed noninvasive, e.g. compared to other experimental variability introduced by user, voice, or word. *Dur* could partly evolve with more stability. The implementation also had small changes in the pitch analysis range.

[12] The different syllabic conditions were interesting because they could serve as controls for the focus of the experiment, evolving emotional prosody, plus they could also explore if syllabic conditions mattered.



Current session status: *Iteration* 0. *Comparison* 5.

Please, select the best solution:

Figure 5.2: Pilot study's aiGA web interface

The experiment was conducted online. Participants were asked to set their speaker to a normal volume level, and to maintain the same settings throughout the experiment, and they received other written instructions.[13] A total of $4 * 10 * 6 = 240$ runs were used in the analysis.

5.3.2 Participant feedback on pilot

Informal feedback from the subjects on their experience indicated overall contentment, although some felt the 10 days x 6 tasks setup was lengthy. Also, some felt that ANGRY was not as distinct as SAD ("could have been angrier"), and one person expressed that the sadness had a "pleading" quality to it. Commenting expressed some worry about consistency, and that perhaps latter runs felt more difficult, i.e. some kind of desensitization effect. Some sounds were felt as reflecting a nicer voice than others, and a suggestion was that rating each sound on a scale of the emotion might be preferable than contrasting two sounds.

5.3.3 Pilot results and discussion

Two sets of results were considered in the analysis; for each considered run r, the aiGA's gene allele values for *Int*, *F0M* and *Dur* for the run's final best (highest-ranking) individual given the resynthesized sound files evaluated by the human judges, as well as for the final evolved synthetic model which extended beyond a user's judgements (but the evolved synthetic solution consisted of two values (μ_r, σ_r) per gene allele). The data set of the best individuals is henceforth called BI, whereas the data set of evolved synthetic models is

[13] A couple of times participants were forced to exit their browsers. When that happened, the participants had been instructed to restart.

called ESM. The below analysis mostly focuses on the values of BI and the means or μ_r values of ESM, but the overall distribution also used σ_r. Plotting and statistics was done with `matlab6.1` in the pilot.

The following analysis clarified that emotions' variables yielded distinct overall prosodic profiles, that aiGA indeed seemed to have evolved emotional prosody, and what the two emotions' averaged prosodic models were in the pilot experiment.

The results representing ANGRY vs. SAD's overall distribution of runs, based on ESM for the individual prosodic variables, are in Figs. 5.3 - 5.4, in $[0, 1]$ range with truncation, i.e. given the encoding scheme described in the section above, where for each word used in the resynthesis, 0.5 reflected the NEUTRAL recording's location in the $[0, 1]$ range. The curves can be seen as representing the overall distribution (μ^*, σ^*) for the variables *Int, F0M, Dur*, respectively, where $\mu^* = \frac{\sum_r \mu_r}{n}$, and $\sigma^* = \sqrt{\frac{\sum_r \sigma_r^2}{n}}$, where n is the number of runs r completed (e.g. for SAD runs $n = 120$, or for monosyllabic $n = 80$). The *pooled standard deviation* σ^* is an estimate of the larger population P (with unseen ANGRY/SAD cases). In contrast, the *sample standard deviation* s is larger, and this difference may be due to the number of runs being quite small.[14] For completeness, histograms over the μ_r sample are also included.[15]

Fig. 5.3 shows that the overall distribution separates emotions, with some overlap for *Int* and *F0M*, but not for *Dur*. As expected, the mean of *Dur* was shorter for ANGRY speech, and longer for SAD speech. For the mean of *Int*, the relative position of emotions to each other was as expected, i.e. ANGRY had higher intensity than SAD, but SAD was at the NEUTRAL original. The mean of *F0M* showed opposite behavior than the majority literature, with slightly decreased near NEUTRAL *F0M* for ANGRY, but increased *F0M* for SAD; Burkhardt and Sendlmeier (2000) also noted raised F0 for this emotion, and hypothesized two sadness levels. In contrast, syllabic types did *not* separate, see Fig. 5.4, and thus, do not seem to make a difference for average behavior.[16]

When resynthesizing words with μ^* values, SAD appeared more distinct than ANGRY,and ANGRY differed mildly from NEUTRAL, although certain words seemed angrier. Better SAD synthetic speech has been noted before (Iriondo, Alías, Melenchón, and Llorca 2004). The ANGRY emotion family, i.e. the spectrum of feelings subsumed by ANGRY, may also be more diverse, and thus vary more.

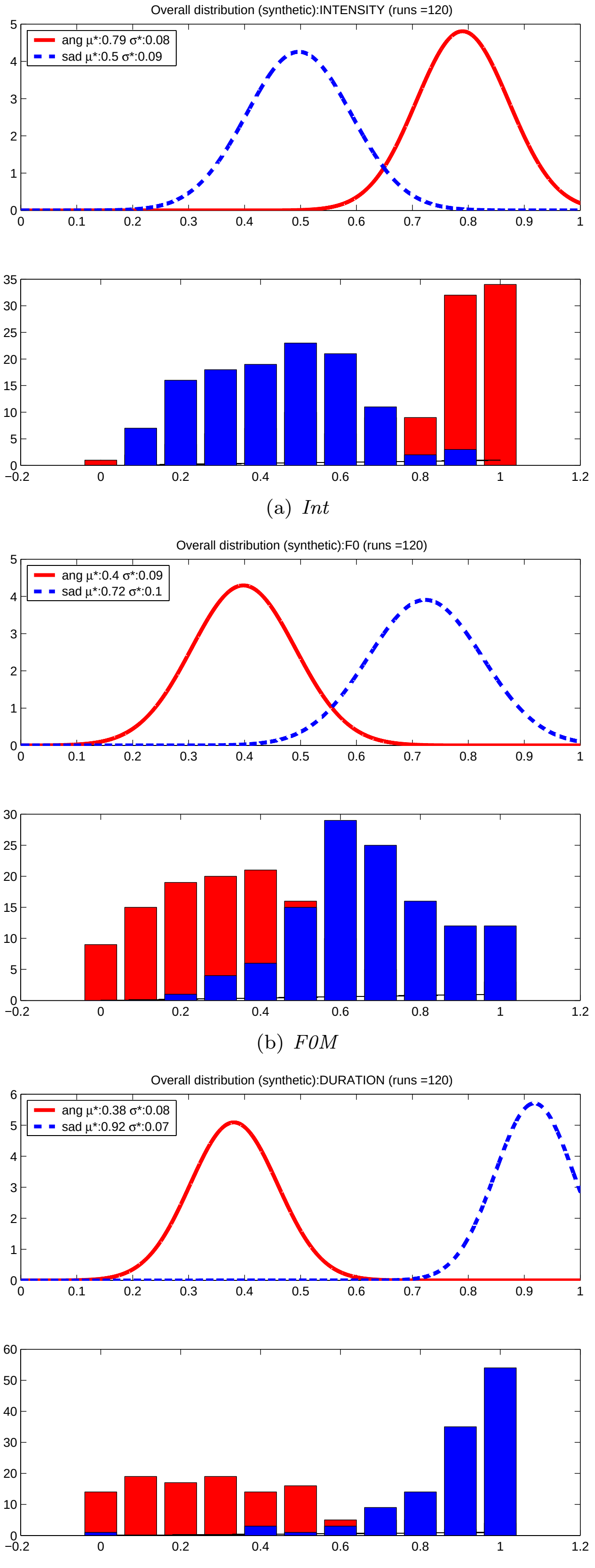
Beyond isolated variables, Fig. 5.5(a-b) visualize runs in 3D as points in encoding in $[0, 1]$ range for three dimensions (*Int*, *F0M*, and *Dur*) for BI and ESM (with truncated μ_r values for ESM).[17] Despite outliers,

[14] The procedure for computing σ^* is valid with the assumption that all runs are samples obtained from a unique population P, that the population variance is unique, and that the parameters of the population remain the same during the runs.

[15] Note that some columns may partially hide others in these histograms.

[16] Similarly, when plotting six curves representing monosyllabic, bisyllabic and trisyllabic word types for the two emotions, these also clustered together by emotion.

[17] Note as caution that 2D and 3D plots are merely descriptive, and may be visually misinforming due to dimensionality, scaling, or point overlap.

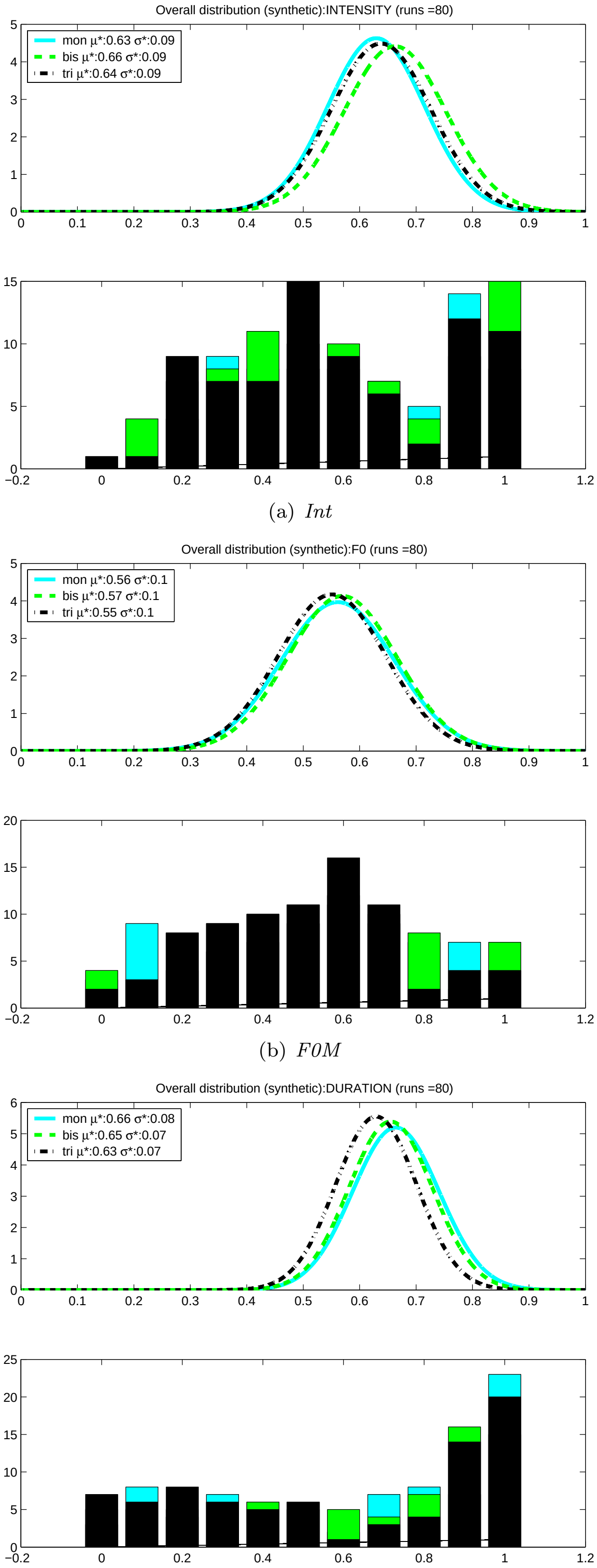


(a) *Int*

(b) *F0M*

(c) *Dur*

Figure 5.3: Overall distributions (μ^*, σ^*) for *Int*, *F0M*, or *Dur* by emotions Sad and Angry show partly or fully separated curves.

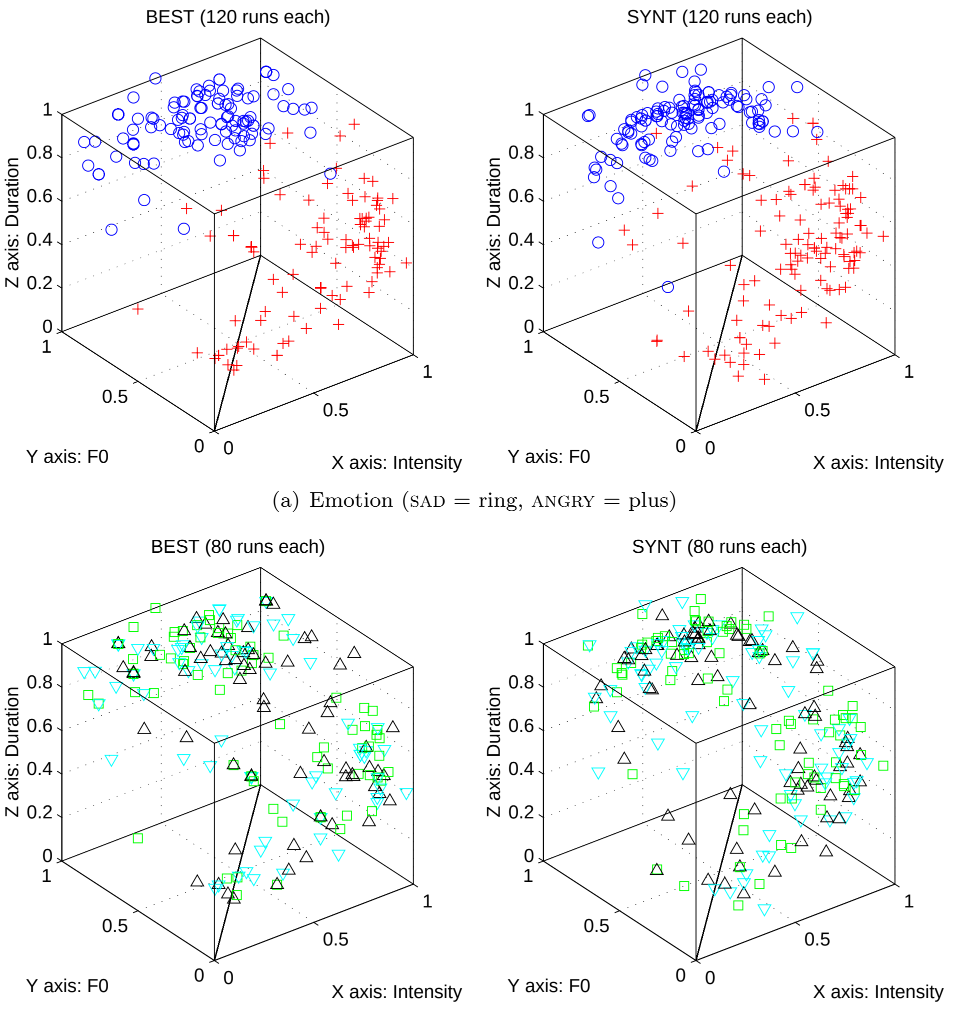


(a) *Int*

(b) *F0M*

(c) *Dur*

Figure 5.4: Overall distribution (μ^*, σ^*) for *Int*, *F0M*, or *Dur* by syllabic type show overlapping curves.



(a) Emotion (SAD = ring, ANGRY = plus)

(b) Syllabic (monos.=↓triangle, bis.=square, tris.=↑triangle)

Figure 5.5: Runs for BI (BEST) vs. ESM (SYNT) in 3D indicate trends for two clouds by emotion (a), but not by syllable type (b).

and quite large estimated s for an emotion given its points and dimensions,[18] Fig. 5.5(a) indicated a trend of 2 clouds of points by emotion, which again contrasted with non-separation by syllabic type in 5.5(b). Although a run's ESM and BI points did not necessarily occur at same place, overall clouds seemed similar for ESM and BI in 5.5(a).[19]

Next, for each prosodic variable, 2-way ANOVAs were done at 95% confidence level for data sets BI and ESM (truncated μ_r), followed by a multiple comparison for significant main factors (using `matlab6.1`'s `anova2` and `multcompare`). Multiple comparison did not consider interactions and should be interpreted with caution. Results are in Table 5.3. The first test considered *syllable types and emotions*, and only the

[18] For example, for BI $s_{sad} = 0.32, s_{ang} = 0.43$ when $s_{emotion_i} = \sqrt{s^2_{Int_{emotion_i}} + s^2_{F0_{emotion_i}} + s^2_{Dur_{emotion_i}}}$

[19] 16% of BI ANGRY equaled the individual set to literature values.

Var.	Fact1	Fact2	#Rep.	Model	Fact1	Fact2	Interac.	`multcompare` diff.
Int	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
Int	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
F0M	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
F0M	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
Dur	syl (3)	em (2)	40	BI		✓		SAD vs. ANG
Dur	syl (3)	em (2)	40	ESM		✓		SAD vs. ANG
Int	user (4)	em (2)	30	BI		✓	✓	SAD vs. ANG
Int	user (4)	em (2)	30	ESM		✓	✓	SAD vs. ANG
F0M	user (4)	em (2)	30	BI	✓	✓	✓	SAD vs. ANG; A vs. BCD
F0M	user (4)	em (2)	30	ESM	✓	✓	✓	SAD vs. ANG; A vs. B
Dur	user (4)	em (2)	30	BI	✓	✓	✓	SAD vs. ANG; AC vs. BD
Dur	user (4)	em (2)	30	ESM	✓	✓	✓	SAD vs. ANG
Int	user (4)	syl-em (6)	10	BI		✓	✓	(same)
Int	user (4)	syl-em (6)	10	ESM		✓		
F0M	user (4)	syl-em (6)	10	BI	✓	✓	✓	
F0M	user (4)	syl-em (6)	10	ESM	✓	✓	✓	
Dur	user (4)	syl-em (6)	10	BI	✓	✓	✓	
Dur	user (4)	syl-em (6)	10	ESM	✓	✓	✓	

Table 5.3: ANOVAs showed that *emotion* was always significant, but *syllabic type* was not. *User (persons A, B, C, D)* was significant for *F0M*, *Dur*, with interactions. The last column shows differences for significant main factors using `multcompare`. ✓ indicates significant p-values (syl = syllabic types, em = emotions, BI = final best individual, ESM = final evolved synthetic model, ANG = angry).

emotion factor showed significant difference. Interactions were not significant, and perceptual differences appeared due to emotion, and not to syllable type. The second test covered *users (persons A, B, C and D) and emotions.* Again, for all variables, emotion was a significant factor. For *F0M* and *Dur* user was also significant, and interaction between factors was always significant. The third test regarded *users and emotion-syllable type task.* The emotion-syllable type task was a significant factor, and so were interactions (except for *Int* in ESM), as were users for, again, *F0M* and *Dur.* Multiple comparisons showed that all tests grouped by emotion, and for the second and third tests, person A, a linguist, was usually involved when *user* was a significant factor. Feedback indicated A decided more analytically; novice users may be less "contaminated" by formal knowledge. However, user impact remained unclear since significant interactions were observed but not well understood, and only 4 users were involved in the pilot study; drawing any conclusions is thus not really possible. In Table 5.4, the user behavior by emotion, based on prosodic variable (truncated μ_r for ESM) and data set, showed complexity. Variation was quite noticeable, but *Dur* appeared less varied for most subjects, at least for SAD.

CART (as implemented by M. Riley) was used on BI and ESM to see how far the binary distinction between SAD and ANGRY models obtained from runs could be learned, and to inspect what features supported prediction. Each example, labeled either SAD or ANGRY, had values as encoded by the aiGA for *Int*, *F0M*, and *Dur* as features (non-truncated μ_r for ESM).[20] Mean precision, recall, and F-score based on 10-fold cross

[20] Only 1 ESM fold had a decision node with value beyond the $[0, 1]$ range. CART experiments were run on Red Hat Linux

BI-ANG	Person A	Person B	Person C	Person D
Int	0.79 (0.21)	0.73 (0.27)	0.8 (0.24)	0.82 (0.22)
F0M	0.55 (0.27)	0.39 (0.25)	0.49 (0.25)	0.25 (0.19)
Dur	0.24 (0.15)	0.4 (0.28)	0.23 (0.14)	0.47 (0.27)
BI-SAD	Person A	Person B	Person C	Person D
Int	0.41 (0.25)	0.58 (0.2)	0.48 (0.21)	0.46 (0.18)
F0M	0.78 (0.18)	0.64 (0.18)	0.63 (0.24)	0.84 (0.15)
Dur	0.92 (0.1)	0.97 (0.05)	0.94 (0.11)	0.89 (0.13)
ESM-ANG	Person A	Person B	Person C	Person D
Int	0.85 (0.2)	0.71 (0.3)	0.84 (0.21)	0.76 (0.26)
F0M	0.43 (0.16)	0.37 (0.23)	0.51 (0.25)	0.29 (0.18)
Dur	0.35 (0.24)	0.39 (0.29)	0.26 (0.14)	0.53 (0.28)
ESM-SAD	Person A	Person B	Person C	Person D
Int	0.44 (0.23)	0.56 (0.14)	0.47 (0.16)	0.51 (0.21)
F0M	0.77 (0.16)	0.65 (0.16)	0.66 (0.17)	0.81 (0.17)
Dur	0.9 (0.1)	0.98 (0.04)	0.94 (0.17)	0.85 (0.18)

Table 5.4: Users' means by emotion for BI and ESM (sample standard deviation in parenthesis; n=30 replicas)

Em-model	Mean prec.	Mean recall	Mean F	% non-unique exs.
ANG-ESM	0.90	0.88	0.88	0.05 (3 types)
ANG-BI	0.95	0.91	0.92	0.28 (7 types)
SAD-ESM	0.90	0.88	0.88	0.05 (3 types)
SAD-BI	0.92	0.94	0.93	0.26 (8 types)

Table 5.5: 10-fold cross validation means from CART classifying SAD and ANGRY evolved synthetic models (ESM) and best individuals (BI). Data set had 240 instances (50% SAD vs. ANGRY), i.e. 24 test examples in each fold.

validation are in Table 5.5.[21] Interestingly, despite the sample variation, on average CART performed well at distinguishing SAD and ANGRY instances. For ESM, 0.9 mean precision, 0.88 mean recall, and 0.88 mean F-score was obtained for both SAD and ANGRY predictions. For BI, performance even increased slightly, which may relate to BI having more repeated feature vectors, see column five in Table 5.5. Inspection of decision trees showed that *Dur* was mostly used as sole predictor. Five ESM folds also used *F0M* for predictions, but *Int* was not used. This may indicate a hierarchy of prosodic feature importance, and that some features may be subject to and show more vs. less variability in this data set.

5.3.4 Concluding summary of the pilot study

Given an initial study of 1-word utterances, aiGA was used to obtain average models of prosodic variables for SAD and ANGRY prosody in interactive resynthesis experiments, with sadness possibly appearing more distinct. Syllabic length appeared less important, which supported continued word-level encoding, although

with some use of an earlier version of python (probably 2.2) for feature vector extraction and vector type determination.

[21] With $\frac{9}{10}$ train vs. $\frac{1}{10}$ test, with a different tenth for test in each fold, using the default parameters (except minor tree cutoff to avoid an overfit).

some words seemed better rendered than others with the values obtained from averaging solutions, and user or word-specific influence remained obscure. Overall, the pilot indicated that aiGA had potential for evolving emotional prosody, and that averages of solutions μ values for *Int*, *F0M* and *Dur* behaved differently for the two emotions. Additionally, *F0M* showed an interesting opposite behavior than expected for SAD and ANGRY. Moreover, 3D plotting indicated trends by emotion, and CART models showed that emotion solutions or best individuals across runs were predictable to quite high degree.

5.4 Extended study

This section covers the extended study, which drew on the lessons learned from the pilot study and additionally extended its scope. In general, the extended study was characterized by substantial increased complexity. It used a larger set of emotional targets, a larger group of participants, an increased set of prosodic parameters, utterances with multiple words, and the aiGA's problem size was more than doubled or tripled respectively in different experiments in terms of number of prosodic gene alleles to evolve.

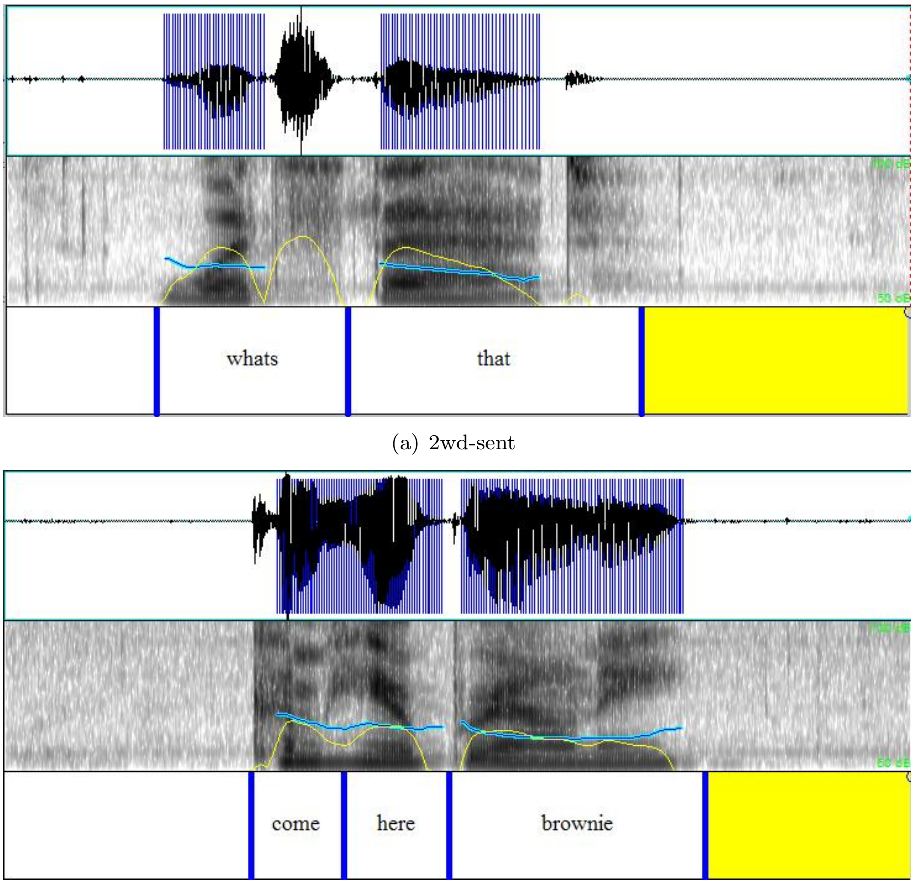
The following first outlines the main differences between the pilot and the extended study, before addressing participants, their perceptions of the study, and finally the empirical results in terms of statistical and graphical analysis, as well as a discussion of resynthesized utterances based on the mean of the evolved aiGA solutions' μ values from different participants for each emotion, before a concluding summary.

5.4.1 Procedure and experimental setup

To begin with, the set of emotions under consideration was extended to five canonical emotions: ANGRY, FEARFUL, HAPPY, SAD, and SURPRISED. Thus, these corresponded to the set of emotions involved in the final text prediction (see Ch. 4).

In the extended experiment, two sentences were paired with each emotion for a particular experimental task (i.e. 5 emotions × 2 sentences = 10 tasks). Intentionally, a single sentential stimuli was used in each experimental task. The manipulated utterance stimuli were longer than in the pilot, namely a two-word sentence, *"What's that?"*, and a three-word sentence, *"Come here Brownie."*, taken from the neutral recordings of story utterances, forming part of the database called *LMB* from Oregon Health and Science University with a female reader.[22] Thus, the sentences represented a gradual increase in size from the pilot's 1-word utterances. Their semantics was flexible enough to fit various emotions, although SURPRISED was arguably less natural for the three-word sentence. The original sentences are henceforth referred to as

[22] Thanks to Esther Klabbers at OHSU who made these recordings available.



(a) 2wd-sent

(b) 3wd-sent

Figure 5.6: Spectrograms of the two-word sentence “What’s that?” and the three-word sentence “Come here Brownie.” with their word splits (the silences at the left and right edges were excluded).

original, and are shown in Figs. 5.6 with their word splits.

For each word, its prosodic parameters were manipulated and resynthesized, and words were then concatenated into an utterance, with Praat used for this process (Boersma and Weenink 2006). The prosodic parameters included intensity (*Int*), duration (*Dur*), as well as pitch (*F0Med*), the latter now reflecting the pitch median in semitones instead of its mean in mel units. The range for pitch analysis spanned from 60 Hz to 515 Hz to cover most post-manipulation. Additionally, a novel fourth parameter, pitch range (*F0Rng*), was added to allow for increasing or decreasing (flatten) the pitch range around the median, or turning the pitch curve upside down (given the Praat version used). Except for *Int*, manipulated individually, the remaining prosodic parameters were modified via Praat’s procedure *Change Gender* in the extended experiment,[23] which employed PSOLA resynthesis with some inherent noise.[24] Praat's *Change Gender* procedure determined pitch modifications according to the following equations from Praat's manual, i.e. the amount of increase or decrease in pitch range was linked to the new pitch median.

$$newPitch = pitch * newPitchMedian/oldPitchMedian \tag{5.1}$$

$$finalPitch = newPitchMedian + (newPitch - newPitchMedian) * pitchRangeScaleFactor \tag{5.2}$$

As in the pilot study, aiGA kept individuals with prosodic parameters for resynthesizing utterances. This time, the variables modified were intensity, pitch median, duration, and pitch range. Thus, each individual again encoded four prosodic gene alleles per word in $[0, 1]$ range, with the *original* at 0.5. In other words, a gene allele value larger than 0.5 represented a relative increase compared to the *original*, whereas a value smaller than 0.5 represented a decrease. However, in this study, a 0.1 step was also set to represent a fixed unit for *Int*, *F0Med*, and *F0Rng*, the size of which depended on whether it was a relative increase or decrease compared to the *original*. Moreover, the maximum ratio for *Dur* was extended. Table 5.6[25] provides ranges and step sizes per prosodic variable. Thus, numerical results should not be explicitly compared to the pilot study, but rather the trends (above, below, and/or around) relative to the *original*.

Variable	Abbreviation	Unit	0.1 step size if allele > 0.5	0.1 step size if allele < 0.5
Sound intensity	*Int*	dB	2.5	-1
Median F0	*F0Med*	semitones	1.5	-1
Pitch range[26]	*F0Rng*	factor	0.4	-0.4
Variable	Abbreviation	Unit	Max boundary	Min boundary
Total duration	*Dur*	ratio	2.1 (longer)	0.7 (shorter)

Table 5.6: Word-level prosodic variables used in the extended study. *Int, F0Med*, and *F0Rng* had defined 0.1 stepsizes, i.e. allowing for 5 steps increase and 5 steps decrease from the *original* at 0.5 given the $[0, 1]$ range, whereas *Dur* was dynamic for each word, given upper and lower bounds.

Since the gene encoding was maintained at the local word-level, the number of prosodic gene alleles representing a resynthesized utterance grew from 3 in the pilot to 8 (2×4) for the two-word sentence and 12 (3×4) for the three-word sentence. Each task involved several pair-wise comparative judgments of sound files; 22 pairs for the two-word sentence, and 46 pairs for the larger three-word problem.

[23] The "pitch range factor" of the *Change Gender* procedure in the used Praat version (probably v. 4.5.01) could modify the pitch range around the pitch median. 2 meant doubling the pitch range, -1 turning the curve upside down (Boersma, personal communication), whereas "0.0 monotonizes the new sound to the new pitch median." (Praat's manual). Formant shift, which could be manipulated for male-female conversion, was kept constant at 1, meaning no change according to Praat's manual.

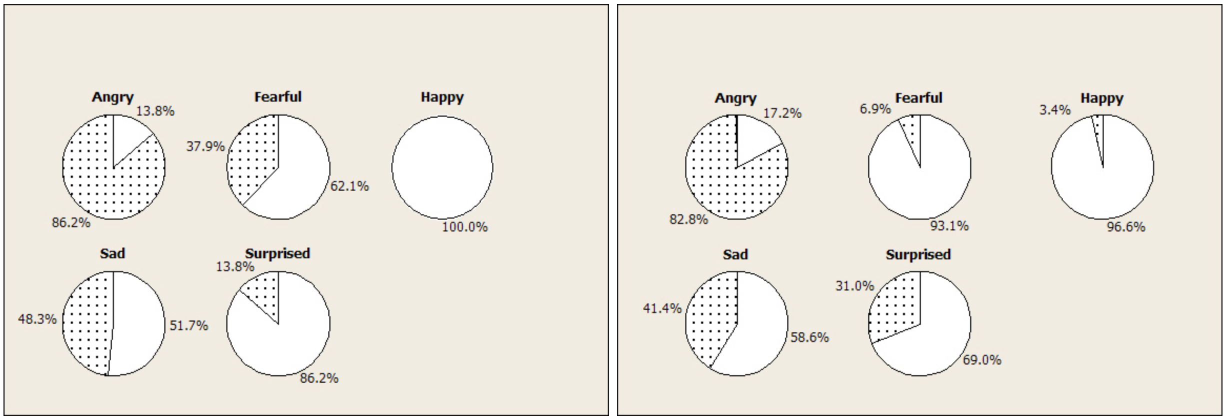
[24] Resynthesis with Praat had some numerical variability from the estimated target (pitch range seemed most affected, and duration least). For example, Praat's PSOLA causes some random variability (Boersma, personal communication), and automated pitch analysis generally involves tracking mistakes. Overall, I expect some inconsistency introduced in the resynthesis, and some formant variation, clipping at extremes, and signal abasement given tentative modification ranges; establishing search space ranges more systematically remains a challenge. A post-experimental release announced a bug fix in *Change Gender* (and discontinued the negative pitch range factor; reason unknown), assumed of low impact given the setup.

[25] These values reflect a developmental version of the Praat script, believed faithful to the experimental code.

As in the pilot, one individual was kept at constant fixed values in each run, and the remaining population was initialized at random. In the extended study, the fixed individual was set manually for each emotion and for each sentence, with the restriction that its values were between 0.15 and 0.85. The fixed settings reflected the experimenter's perceptions, although for SAD and ANGRY, trends from the pilot were considered. Each task's fixed individual can be regarded as a rather quick attempt at establishing a hypothesized near solution with the intent to help speed up the search. Values for the fixed individuals are in Table 5.7, and the ratio of fixed individual finishing as top-ranked best individual in the experimental data is shown in Figs. 5.7, with the largest ratios occurring for ANGRY, followed by SAD, i.e. the emotions explored in the pilot. Differences in the ratios of fixed individuals indicated that their quality differed by emotion.

Emotion	I_{w1}	FM_{w1}	D_{w1}	FR_{w1}	I_{w2}	FM_{w2}	D_{w2}	FR_{w2}	I_{w3}	FM_{w3}	D_{w3}	FR_{w3}
$angry_{2wd}$	0.85	0.2	0.15	0.15	0.85	0.2	0.15	0.85	-	-	-	-
$angry_{3wd}$	0.85	0.2	0.4	0.4	0.79	0.4	0.4	0.4	0.79	0.15	0.5	0.2
$fearful_{2wd}$	0.4	0.8	0.3	0.8	0.85	0.85	0.3	0.8	-	-	-	-
$fearful_{3wd}$	0.85	0.15	0.2	0.5	0.85	0.85	0.8	0.6	0.6	0.7	0.2	0.5
$happy_{2wd}$	-	-	-	-	-	-	-	-	-	-	-	-
$happy_{3wd}$	0.85	0.7	0.85	0.4	0.79	0.4	0.4	0.4	0.7	0.4	0.85	0.7
sad_{2wd}	0.5	0.72	0.85	0.25	0.5	0.72	0.85	0.4	-	-	-	-
sad_{3wd}	0.5	0.72	0.85	0.25	0.5	0.72	0.85	0.25	0.5	0.72	0.85	0.3
$surprised_{2wd}$	0.8	0.6	0.3	0.85	0.8	0.85	0.3	0.15	-	-	-	-
$surprised_{3wd}$	0.85	0.7	0.3	0.7	0.2	0.6	0.2	0.2	0.6	0.7	0.15	0.4

Table 5.7: Fixed individual's values for prosodic gene alleles by emotion and sentence ($2wd$ = two-word sentence, $3wd$ = three-word sentence, I = *Int*, FM= *F0Med*, D = *Dur*, FR = *F0Rng*, $w1$ = first word, $w2$ = second word, $w3$ = third word). (The values used for the two-word HAPPY were not verifiable.)



(a) 2wd-sent

(b) 3wd-sent

Figure 5.7: The ratio of the fixed individual finishing as top-ranked best individual in 29 runs is dotted, and was highest for ANGRY, followed by SAD, for both sentences.

The experiment took place in a room in the NCSA building on University of Illinois' Urbana campus, and lasted around 2-3 hours depending on the session's group size.[27] The requirements for participation

[27] The building's A/C meant some background noise (not commented on in the postsurvey).

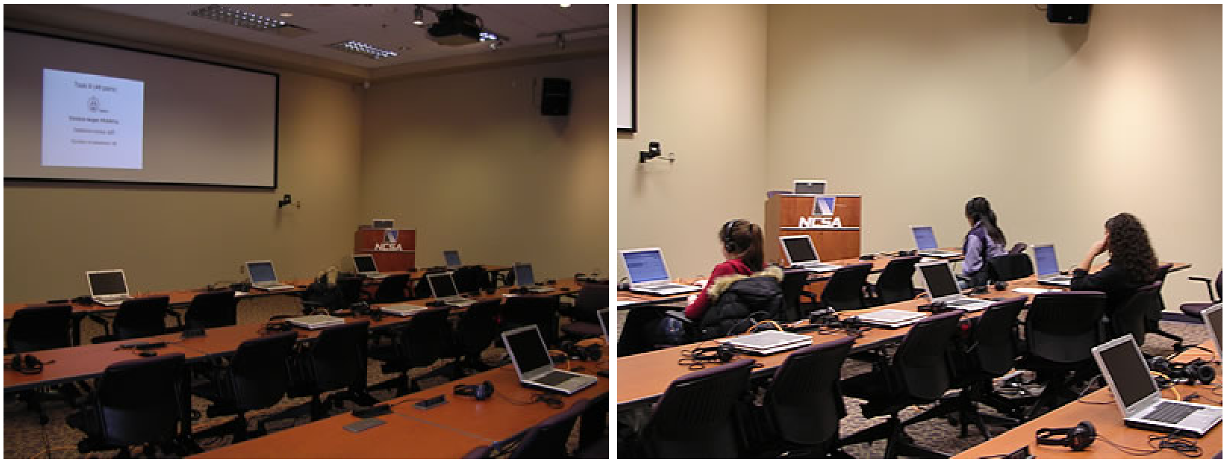


Figure 5.8: Experimental room setup

were being a US native English speaker, US citizen, at least 18 years old, with normal hearing. Each participant was seated at a Fedora Linux laptop computer, as shown in Figs. 5.8, and used Sony MDR-V150 headphones. The subjects were paid for their participation. First, participants read instructions (see appendix B), followed by a short presentation repeating main points. Next, the participants conducted the five tasks for the two-word sentence,[28] followed by a 10-15 minute break with cookies and refreshments, which intended to decrease fatigue, before the final five tasks with the three-word sentence. A lottery had determined a fixed order of emotion tasks as well as each emotion task's randomization seed for the aiGA (same for each participant and sentence).[29] Between each task, the participants also had a very brief break as the experimenter reset the laptop for the next task. Finally, after finishing all tasks, the participants completed a short survey (see appendix C), which is discussed next.

5.4.2 Survey responses

Most survey questions were unrestricted and open-ended to allow for qualitative complementary insights into the study. The following does not claim to cover all responses given, but rather intends to briefly summarize observations across most or part of the participant body, or selected particularly interesting parts of their comments. Note that [sic!] indicates an orthographic mistake in a particular survey response.

Characterization of participant body

22 females and 7 males participated in the study. The participants' age ranged from 18 – 22. Participants reported coming from a wide variety of majors or *Undecided*. Five students studied Linguistics or a

[28] Each task was announced on a separate slide. The main interface had some modifications compared to the pilot, such as a RealPlayer plug-in for easier clicking. The written instructions encouraged discriminate judgements, compared to ties.

[29] The order was FEARFUL, HAPPY, SURPRISED, SAD, ANGRY, and seeds were taken from among those used in the pilot.

particular language area (French or Spanish). Other majors included, for instance, Aerospace engineering, Civil engineering, Early childhood education, Psychology, History, and Advertising. The average number of courses attended which related to linguistics or the study of language was 0.7 (ranging from 0 to 4).[30]

Perceptions of the study

On the one hand, many participants commented positively on the setup of the study; e.g. that the experiment was "very well-run, consistent, and professional. The program was very straightforward too", "very well organized & efficient", "ok", "impressive", "well set up + gave great directions", or that "[t]he experimenter was very nice. I enjoyed this experiment compared to past experiences", "I would do it again", and so on.[31]

On the other hand, as one would expect for this sort of recurring experimental tasks, a fairly frequent comment was that it was e.g. "repetitive", "lengthy", "very monotonus[sic!]", "a little long", "slightly boring", "really tired of the phrase 'come here brownie' [...] :)", etc. Nevertheless, comments included that it, e.g. "wasn't that bad", "painless and went by rather quickly." Moreover, perceptions of boredom or discontent did not necessarily seem to mean disliking the experiment. In fact, the experiment was rather frequently characterized as "interesting", and a few showed curiosity or hypothesized briefly about the experimental outcome or motivations.

Suggestions included, for example, that a larger number of different sentences may have improved the experience. One person commented on getting fatigued "towards the end."[32] Another person said that she "appreciated the breaks between tasks because it helped me wake up & be alert."

Perceptions of emotional targets

The survey responses indicated that many, although not all, participants found certain emotions more or less satisfactorily rendered, but also that individuals' perceptions were quite different and sometimes contradictory. Interestingly, *all emotion targets* were mentioned as being better rendered/easier to discern or identify, as well as being worse/more difficult. In other words, variation appeared to exist among individuals' preferences and perception of emotional speech in this experiment.[33] However, some emotions were more often mentioned than others in certain contexts. For example, SURPRISED was most frequently mentioned as difficult or unsatisfactory (by around one third of participants) followed by FEARFUL and HAPPY, whereas

[30] Basic language program courses, e.g. German 101-104, were not counted.

[31] One person said it was "a bit confusing but I bet it could produce some unique results".

[32] The experimenter noticed that the experiment could be tiring. However, a few already appeared sleepy when entering the experiment, and one person apologized for having been "sleepy".

[33] This relates to question 4 in the survey, see appendix C. Since participants here were not informed about the underlying evolutionary algorithm, perhaps a more direct formulation would have been to ask if some emotions were easier to perform pair-wise judgements for, and if they felt certain characteristics specified particularly well-rendered sound files for each of the emotions.

ANGRY was most frequently mentioned as satisfactory (by more than one half of the subjects) followed by HAPPY and SAD. A few also felt that a given sentence better portrayed one or more emotions.

Critique voiced by some participants included feeling, e.g. that sound files did not fit targets, that certain emotions had less suitable solutions, that differentiating could be hard, or the challenge in separating meaning and voice. One participant noted that "[i]t was had[sic!] to identify surprise when the statement wasn't a question." In terms of strategies for selecting solutions, one person commented that "[t]hey[34] all sounded the same but the emotion target influence whether I choose one target [solution] over another", whereas another strategy described was using mental imagination.

Additionally, my impression was that describing prosodic characteristics was challenging for participants who attempted to do so. Such comments tended to be short, quite generic, and describe one or two emotions, e.g. "[a]nger, sounded more flat whereas happy was portrayed by holding the vowel sounds longer & saying the phrase in a more high pitched voice", "whether something (a syllable) was drawn out or short often helped me to categorize some of the sound files", "drawing out words made [sic!] to work. Upward inflections added to it as well", "[anger:] a certain sterness[sic!] in the voice", "[angry and happy:] you look for if the pitch goes higher or lower at the end of a sentence", "voices were higher for happy and lower for angry", "[s]ad ones always sounded slower with a more decreasing tone. Surprised and angry were quicker", "certain words had more emphasis & certain tracks were louder or longer/shorter", or "[s]urprise meant happy energetic or sudden shift in voice pattern". These responses indicate a dominant laymanship among the participant body, with lack of formal knowledge of prosody.

Lastly, the following reflect a few selected negative and positive perceptions: "some of the recordings were difficult to understand", "for most of them, they weren't too close to how I would speak", "some were very hard to tell and many seemed very computer generated"; versus "[i]nteresting because slight variations in the solutions completely changed my perceptions of the moods conveyed", "[i]t was a good indicator of different tonal inflections", "I could tell the differences in voice intonation easily."

5.4.3 Empirical results

This part describes the empirical results of the extended aiGA study introduced above. A statistical analysis is followed by a graphical analysis. Each analysis used the μ values for individual prosodic gene alleles for individual words from resulting evolved synthetic models of different participants' runs.[35]

[34] Ambiguous whether she meant all or specific emotion targets.

[35] The aiGA solutions were in $[0, 1]$ range, except two values were found that went beyond 1 in the aiGA's output of final evolved models.

ANOVAs

Using MINITAB, one-way ANOVAs were conducted for the five-level factor *emotion* (ANGRY, FEARFUL, HAPPY, SAD, SURPRISED). A separate test was conducted for each prosodic gene allele, i.e. for each prosodic variable separately for each word, given their resulting μ values of the evolved synthetic solutions obtained from all 29 runs, each involving one participant. As shown in Tables 5.8 and 5.9, for both the two and three-word sentences all tests were significant, and all except one at $p \leqslant 0.01$.

Prosodic variable$_{word}$	*Emotion* as significant factor for this prosodic gene allele?
Int_{w1}	✓
Int_{w2}	✓
$F0Med_{w1}$	✓
$F0Med_{w2}$	✓
Dur_{w1}	✓
Dur_{w2}	✓
$F0Rng_{w1}$	✓
$F0Rng_{w2}$	✓

Table 5.8: Results for two-word sentence of one-way ANOVAs for factor *emotion* with five levels: ANGRY, FEARFUL, HAPPY, SAD, SURPRISED. Each prosodic variable was subjected to individual tests per word, based on solutions' μ values (29 runs, one per participant). All tests were significant at $p \leqslant 0.01$.

Prosodic variable$_{word}$	*Emotion* as significant factor for this prosodic gene allele?
Int_{w1}	✓
Int_{w2}	✓
Int_{w3}	✓
$F0Med_{w1}$	✓
$F0Med_{w2}$	✓
$F0Med_{w3}$	✓
Dur_{w1}	✓
Dur_{w2}	✓
Dur_{w3}	✓
$F0Rng_{w1}$	✓
$F0Rng_{w2}$	✓
$F0Rng_{w3}$	✓

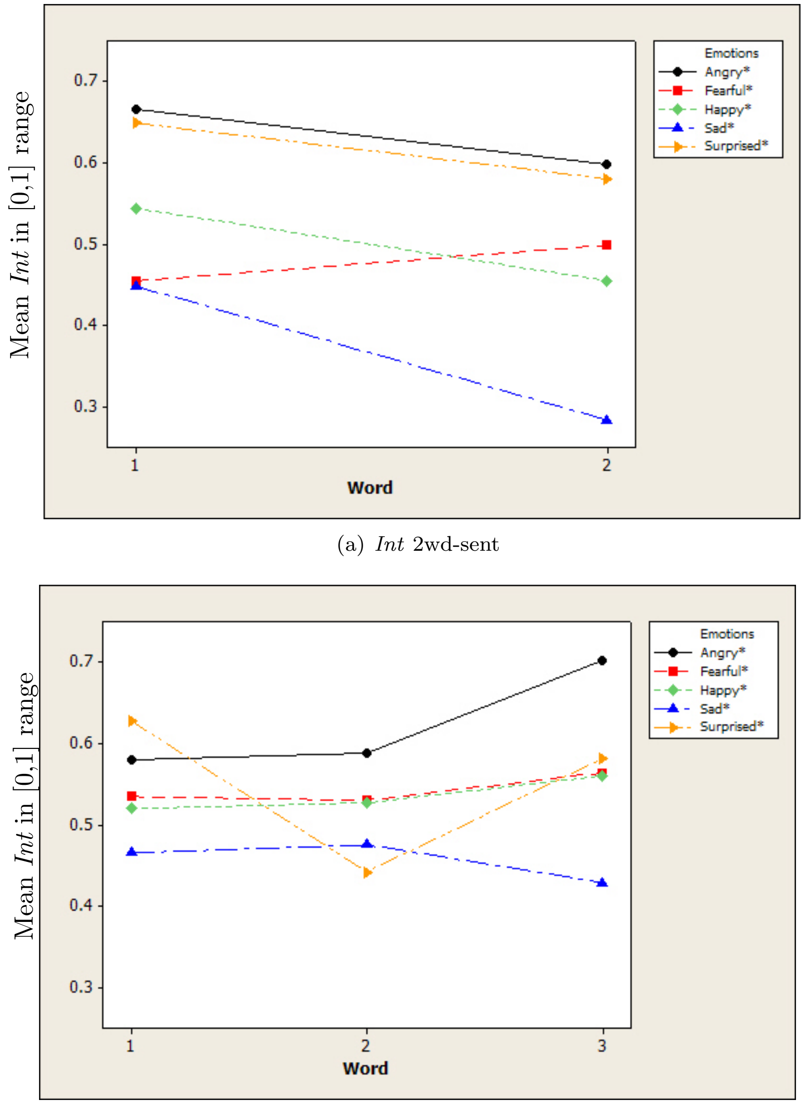
Table 5.9: Results for three-word sentence of one-way ANOVAs for factor *emotion* with five levels: ANGRY, FEARFUL, HAPPY, SAD, SURPRISED. Each prosodic variable was subjected to individual tests per word, based on solutions' μ values (29 runs, one per participant). All tests were significant, all except Dur_{w2} at $p \leqslant 0.01$.

Thus, the ANOVAs demonstrate that all prosodic variables significantly differed by the factor emotion for each word in both utterances. The next subsection analyzes the prosodic characteristics of each emotion.

Graphical analysis

As mentioned above, averaged results across runs are interesting to examine general trends. Figs. 5.9 - 5.12 show each of the five emotions' means of the μ values of 29 solutions for various words in both sentences, whereas the boxplots in Figs. 5.13 - 5.16 display the ranges of the μ values for the same contexts.

Interestingly, although not examined with statistical tests, descriptively the picture that emerged seemed characterized by a great deal of dynamics, with variation both by word and by sentence for most prosodic variables. The intricate variations at the local and inter-utterance or inter-word levels were by themselves an interesting finding, i.e. going beyond analyzing emotional speech at the global level of the complete utterance. Nevertheless, some trends seemed to emerge, discussed for each prosodic variable individually.

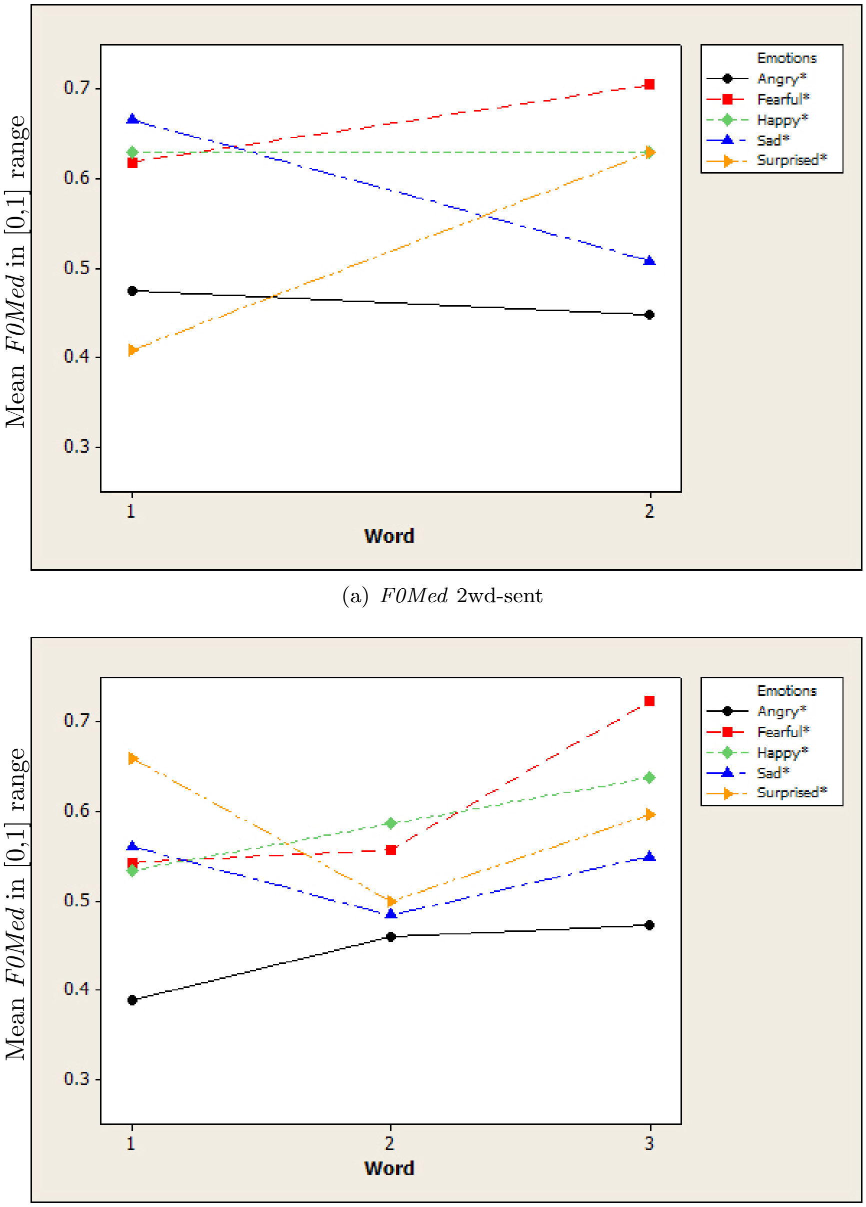


(a) *Int* 2wd-sent

(b) *Int* 3wd-sent

Figure 5.9: *Int*: mean of solutions (μ values) by word ($0.5 =$ *original*), given 29 runs

Although the placement and directions of the curves for *Int* in Figs. 5.9(a-b) differed by sentence, the relative location of emotions patterned with respect to each other. On average, ANGRY tended to have higher *Int* (above *original*) vs. lower for SAD (below *original*). FEARFUL and HAPPY tended mostly to be located in-between, whereas SURPRISED might fluctuate, as especially noticed for the three-word sentence.



(a) *F0Med* 2wd-sent

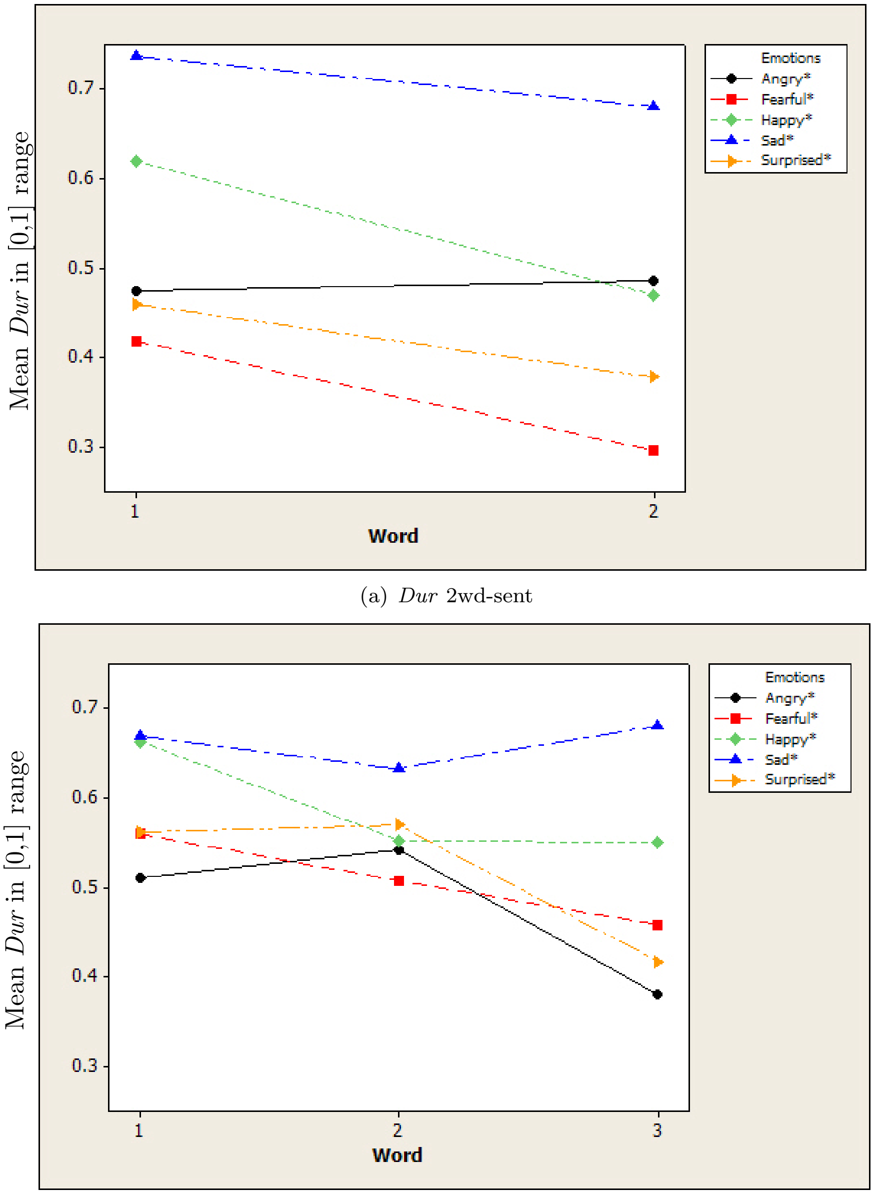
(b) *F0Med* 3wd-sent

Figure 5.10: *F0Med*: mean of solutions (μ values) by word (0.5 = *original*), given 29 runs

For averaged *F0Med* and *Dur* some patterns emerged across the sentences, although the data overall appeared less clear as seen in Figs. 5.10(a-b) - 5.11(a-b).

As shown in Figs. 5.10(a-b), disregarding SURPRISED which again seemed marked by fluctuation, the average of ANGRY's μ values had the lowest *F0Med* (below *original*), whereas FEARFUL, HAPPY and SAD took on slightly higher values for *F0Med*. It is also interesting to notice that the three latter emotions in both cases had grouped averages for the initial word which then diverged.[36]

[36] This cannot simply be explained by the values of the fixed individual as those differed, see Table 5.7.



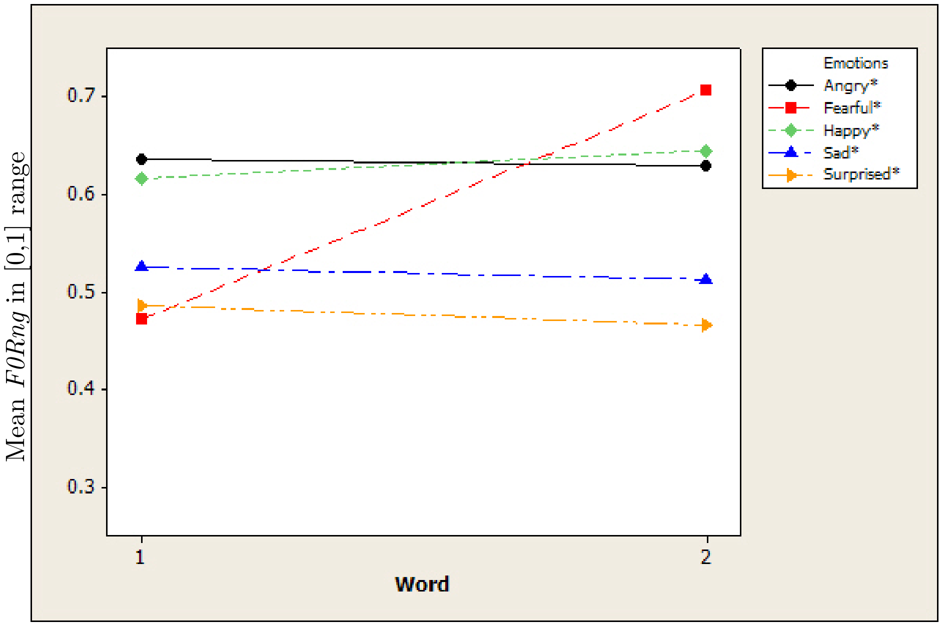
(a) *Dur* 2wd-sent

(b) *Dur* 3wd-sent

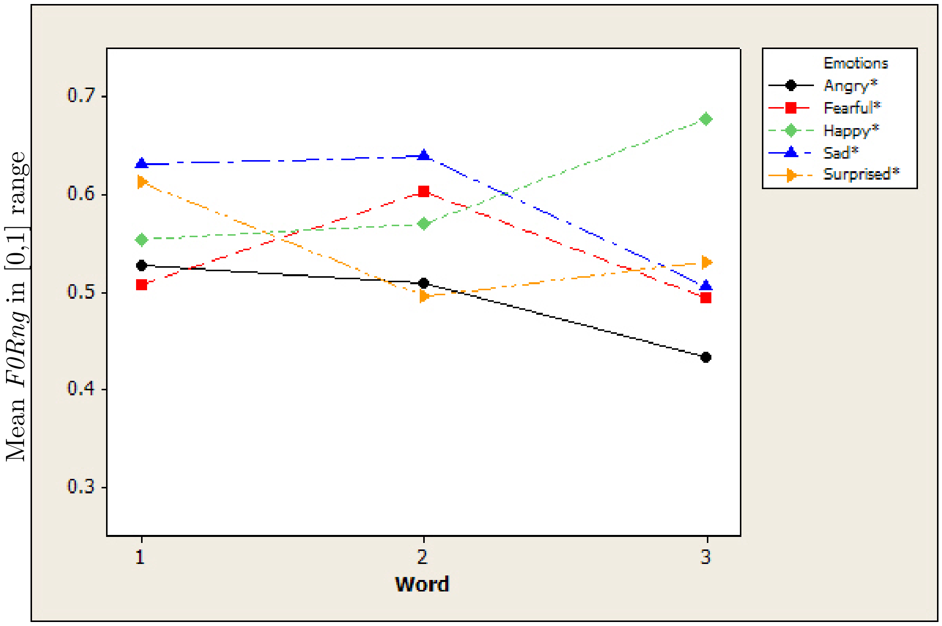
Figure 5.11: *Dur*: mean of solutions (μ values) by word ($0.5 =$ *original*), given 29 runs

Figs. 5.11(a-b) shows that in the case of *Dur*, SAD showed longer word *Dur*, whereas HAPPY and FEARFUL both showed a tendency for a shorter word *Dur* after the initial word.

Lastly, on average, Figs. 5.12 reveal that *F0Rng* did not seem to display a clear trend, except that HAPPY consistently showed a somewhat increased pitch range compared to the *original*. A flattened or turned-up-side-down pitch curve was generally not supported. The lack of a trend could be related to *F0Rng* depending on *F0Med*. Resynthesis inconsistencies could also affect convergence.



(a) *F0Rng* 2wd-sent



(b) *F0Rng* 3wd-sent

Figure 5.12: *F0Rng*: mean of solutions (μ values) by word (0.5 = *original*), given 29 runs

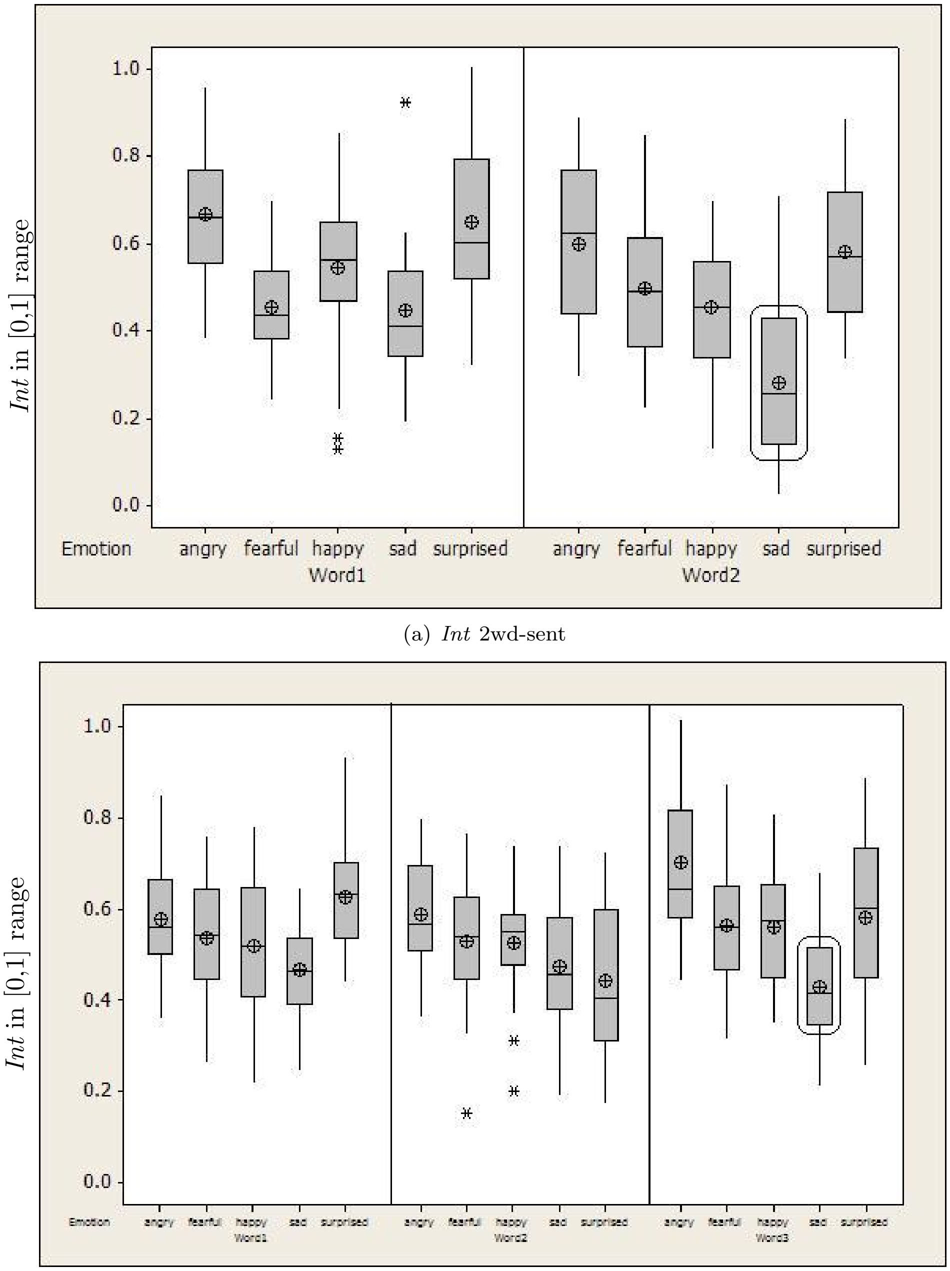
Emotion	Int	F0Med	Dur	F0Rng
angry	$>$	$\leqslant$		
fearful		$\geqslant$		
happy		$\geqslant$		$>$
sad	$\leqslant$		$>$	$\geqslant$
surprised				

Table 5.10: Emotions' approximate average directional behavior of increase or decrease compared to *original* for prosodic variables across words, given the mean of μ values from 29 runs' solutions. The averaged data ranged from approximately 0.7 - 0.3 with *original* at 0.5. Empty cells signifies that average of μ values both exceeded and fell below the *original* across words and sentences, as seen in Figs. 5.9 - 5.12.

Thus, Table 5.10 summarizes the approximate tendencies across words and sentences for emotions compared against *original* at 0.5 in terms increase or decrease, given the mean of prosodic gene alleles' evolved μ values from 29 solutions. Empty table cells indicate that means both exceeded and fell below *original*. Means did not reflect extreme values, and a prosodic variable's range for a given emotion often included the *original*, except for *Int* for ANGRY and *Dur* for SAD. That ranges occurred for prosodic variables might hint at that fluctuation across words itself might play an important role in emotional speech. It may also reflect a leveling effect due to a larger participant group, and/or that non-extreme emotional speech was preferable for the participant group.

Moreover, some local trends may have occurred across sentences, as circled in Figs. 5.13 - 5.15. To begin with, SAD's mean and interquartile range for *Int* were somewhat lower on the final word; more so for the two-word sentence. Similarly, FEARFUL increased its *F0Med* such that the highest mean and interquartile range for *F0Med* occurred on the final word. Additionally, for HAPPY and FEARFUL the means appeared somewhat higher for *Dur* initially. Another interesting observation from the boxplots was that for both sentences, the rough pattern between emotions for *Int*, *F0Med*, and *Dur* on the initial word was sometimes repeated on the final word.

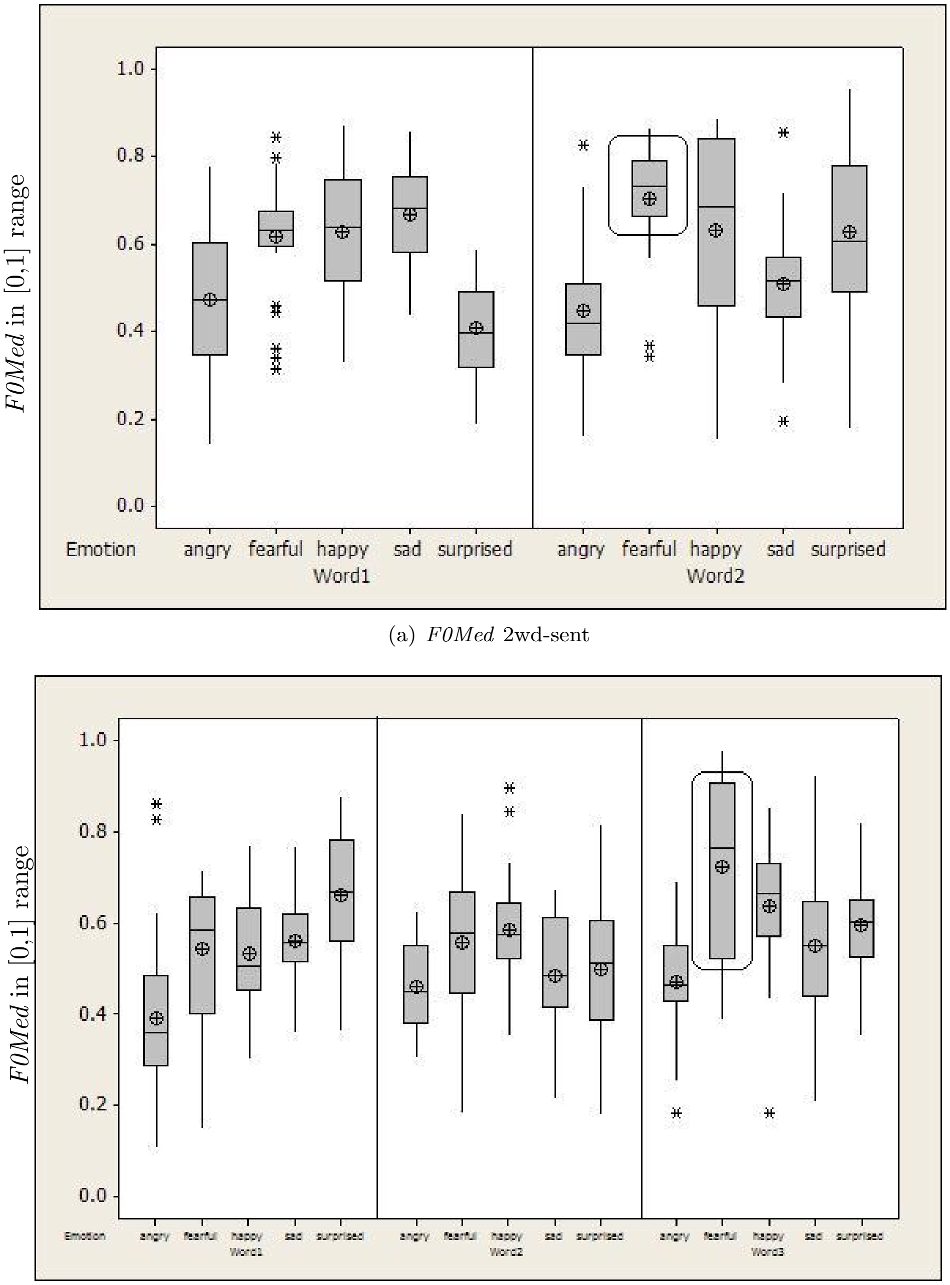
Next, I go on to juxtapose two pairs of emotions in more details, limiting the analysis to the more informative first three prosodic variables. Of the two pairs of emotions examined, one suggests reoccurring rather clear trends, whereas the other is marked by more diversity.



(a) *Int* 2wd-sent

(b) *Int* 3wd-sent

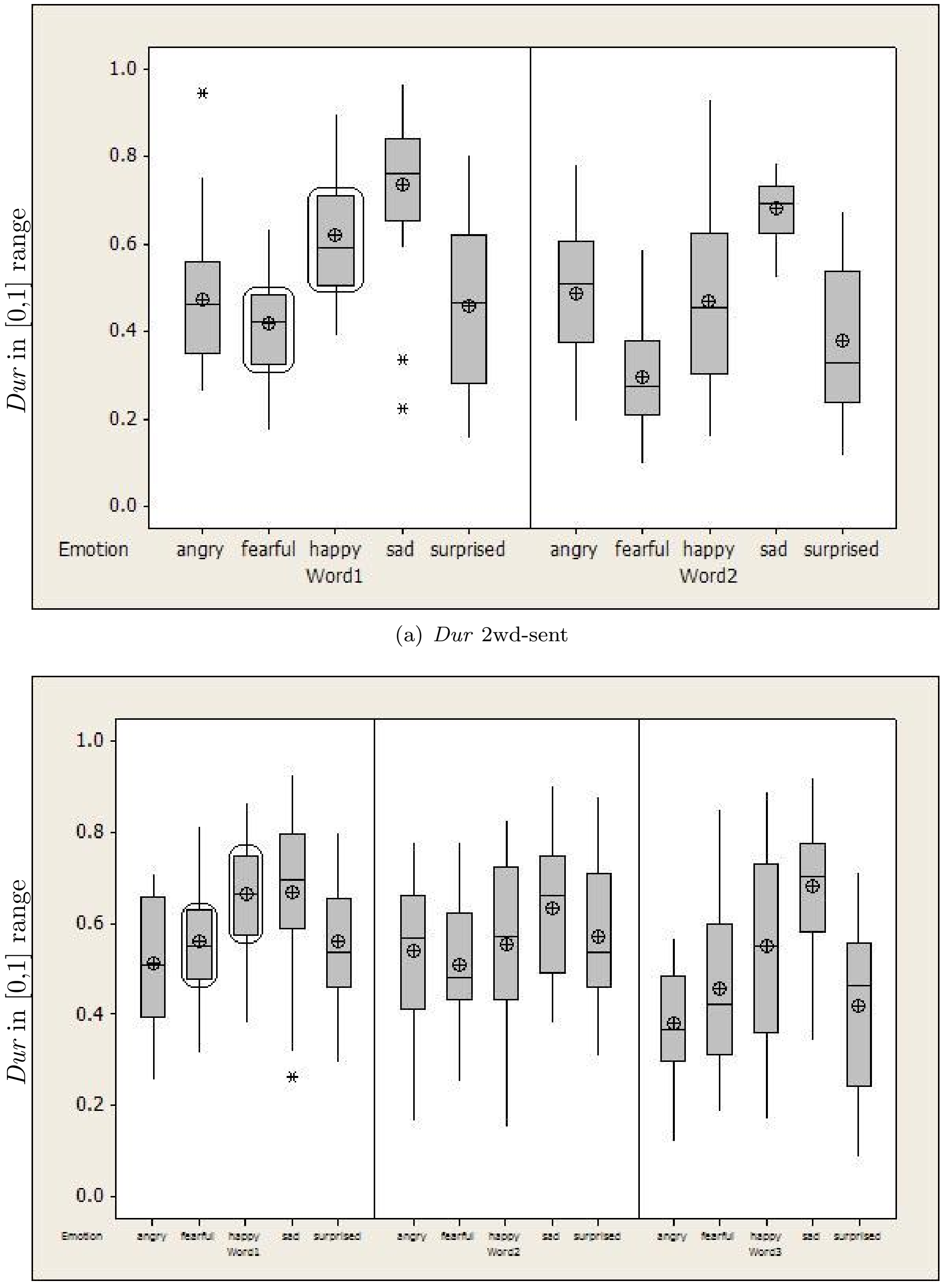
Figure 5.13: *Int*: Spread of solutions (μ values) by word (0.5 = *original*), given 29 runs



(a) *F0Med* 2wd-sent

(b) *F0Med* 3wd-sent

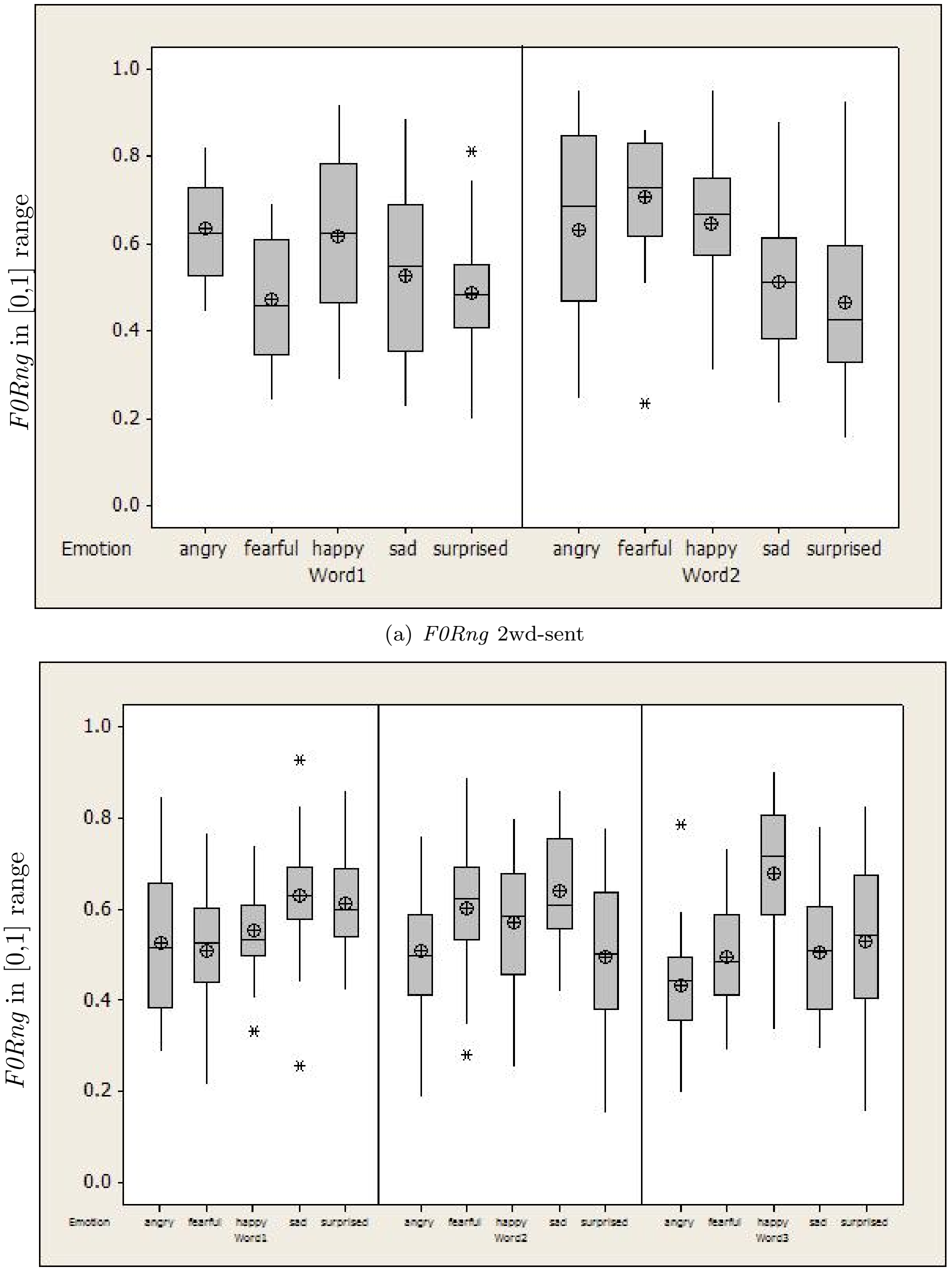
Figure 5.14: *F0Med*: Spread of solutions (μ values) by word (0.5 = *original*), given 29 runs



(a) *Dur* 2wd-sent

(b) *Dur* 3wd-sent

Figure 5.15: *Dur*: Spread of solutions (μ values) by word ($0.5 = \textit{original}$), given 29 runs



(a) *F0Rng* 2wd-sent

(b) *F0Rng* 3wd-sent

Figure 5.16: *F0Rng*: Spread of solutions (μ values) by word (0.5 = *original*), given 29 runs

Case study: SAD vs. ANGRY

If isolating the analysis to the two emotions treated in the pilot study, and looking at the general trends in terms of the emotions' positioning relative to one another for prosodic variables (rather than at specific numerical values) it is clear that the extended study mostly replicated the directional trends for SAD and ANGRY with respect to each other, and now for larger sentences.

Generally, the relative ordering of these two emotions' prosodic variables with respect to each other reflects the trends found in the pilot study, as shown in Figs. 5.17 - 5.18 and the differences in means were mostly confirmed statistically by one-way ANOVAs considering these two emotions as a two-level factor for each individual word.[37] Again, one found higher *Int* for ANGRY than for SAD, as well as longer *Dur* for SAD than for ANGRY. Adding to previous information, *F0Med* now showed some local flavor. Similar to the pilot, *F0Med* was somewhat higher for SAD than for ANGRY initially for both sentences and finally for the three-word sentence, however a difference was statistically not confirmed for the second word for either sentence.[38] Moreover, the boxplots descriptively revealed that there were local differences in spread, interquartile range, mean or median by individual words and sentences.

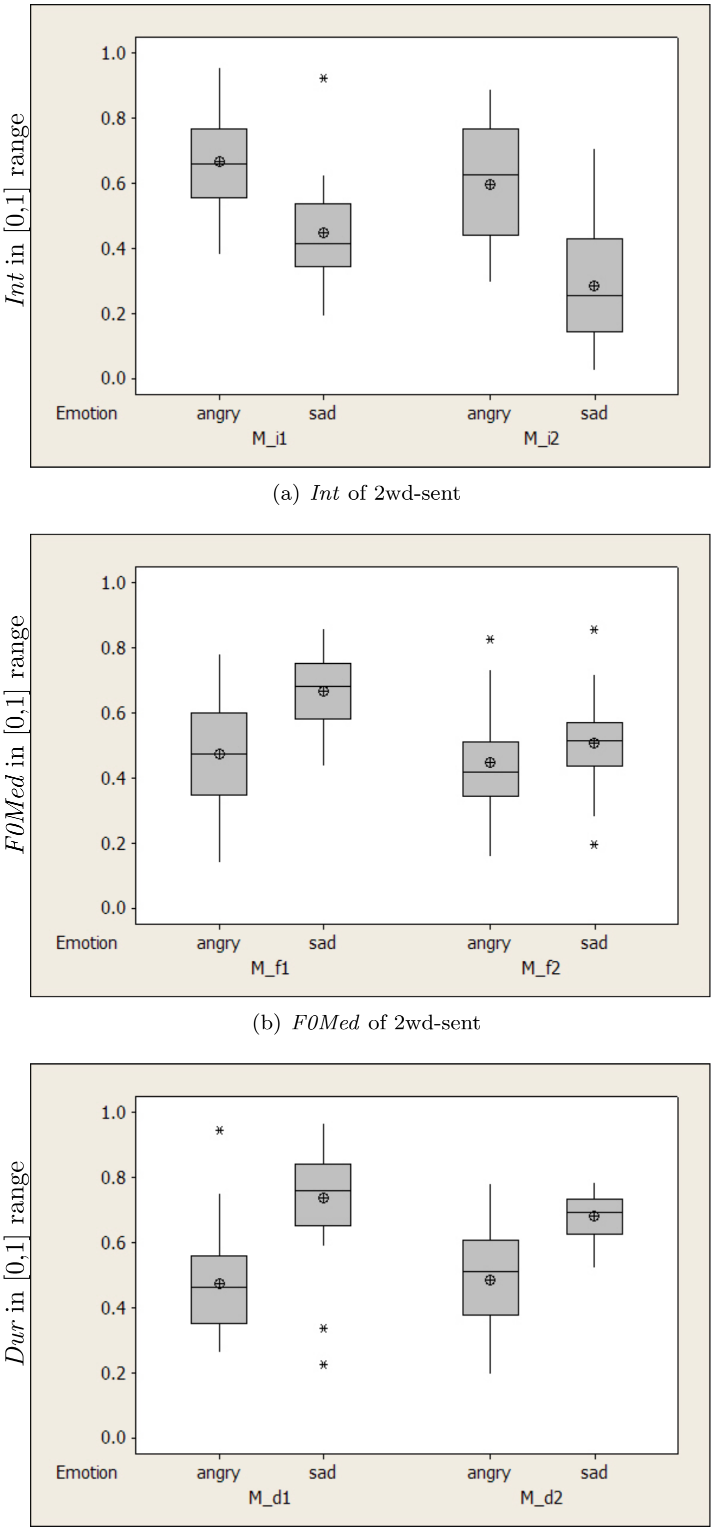
Thus, these largely replicated trends from the pilot under a new and larger participant body seemed to indicate stability of the aiGA approach. It is also possible that since the hypothesized near solution (i.e. fixed individual) for these emotional targets were inspired by the empirical results from the earlier pilot study, their search may have had a stronger setup. Moreover, as noted in Sec. 5.4.2 above, several participants commented on especially ANGRY but also SAD reflecting more satisfactory renderings or being easier targets, which lends support for the averaged directional trends for these two emotions.

Case study: FEARFUL vs. HAPPY

Compared to the fairly clear case of ANGRY VS. SAD, juxtaposing FEARFUL and HAPPY in Figs. 5.19 - 5.20 and by one-way ANOVAs revealed a more complicated picture, with more diversity by sentence. For the two-word sentence, the differences for *Int* was significant for the first word, and the boxplot shows that it was higher for HAPPY compared to FEARFUL. But this did not hold for its second word nor for any word of the three-word sentence. Differences for *F0Med* was not significant in both sentences, although means were above *original* for the two-word sentence. However, the boxplots also indicated that both sentences had in common that the final word was marked by a local increase in *F0Med* for FEARFUL. *Dur* was significant for each word for the two-word sentence, and the boxplot showed that FEARFUL consistently had significantly

[37] I continued using the ANOVA (rather than a t-test) since it had been applied to the five-level scenario above.

[38] *F0Rng* was also statistically significant, however showed conflicting directions for the two sentences.

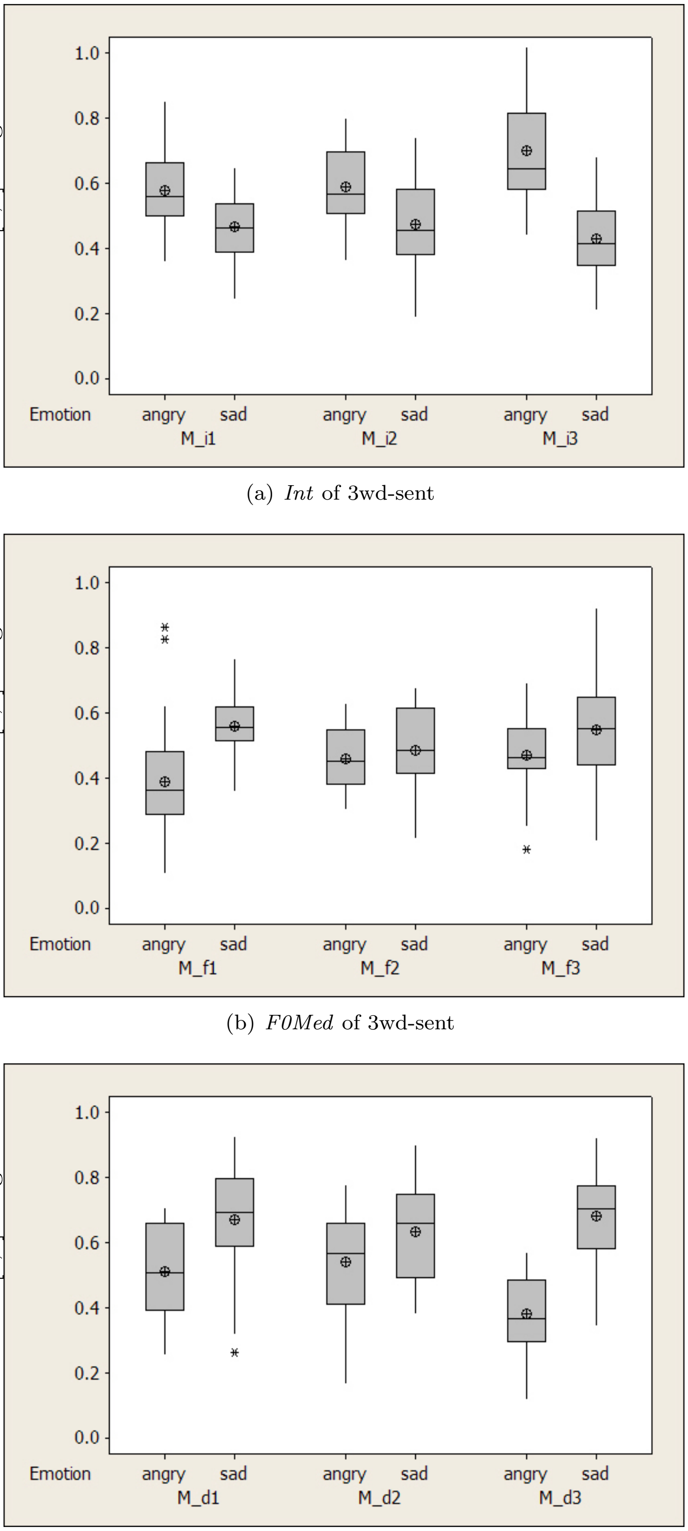


(a) *Int* of 2wd-sent

(b) *F0Med* of 2wd-sent

(c) *Dur* of 2wd-sent

Figure 5.17: ANGRY vs. SAD: Evolved prosodic solutions for 2-word sentence from 29 runs each (μ values)

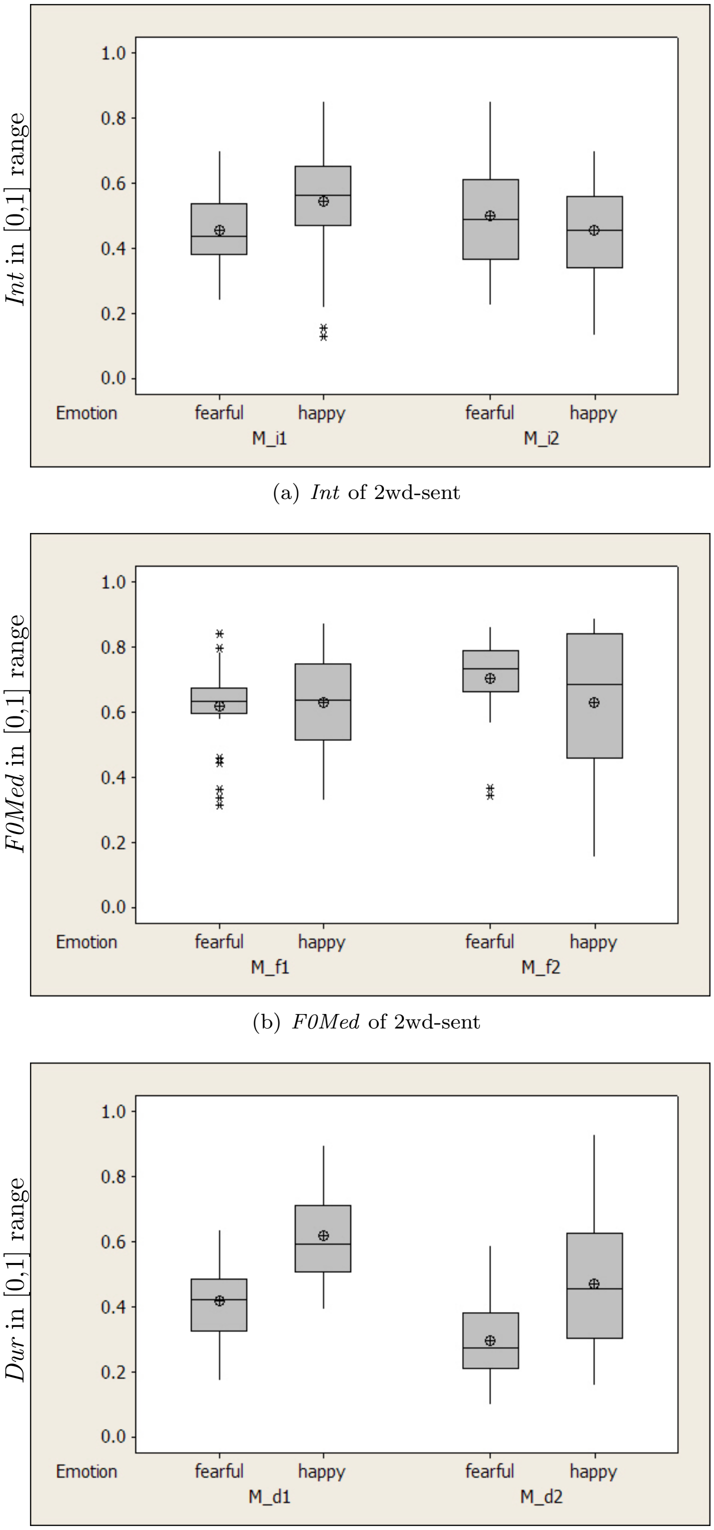


(a) *Int* of 3wd-sent

(b) *F0Med* of 3wd-sent

(c) *Dur* of 3wd-sent

Figure 5.18: ANGRY vs. SAD: Evolved prosodic solutions for 3-word sentence from 29 runs each (μ values)

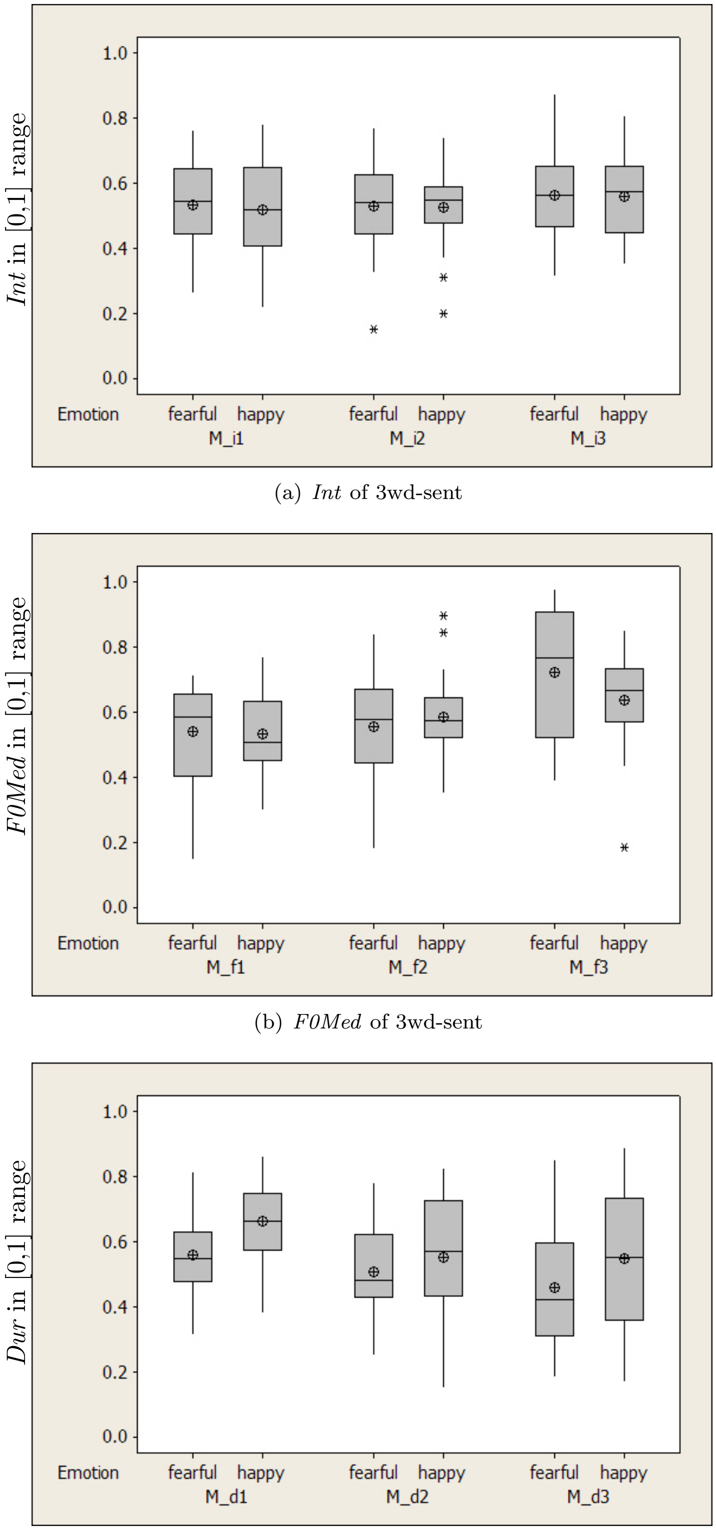


(a) *Int* of 2wd-sent

(b) *F0Med* of 2wd-sent

(c) *Dur* of 2wd-sent

Figure 5.19: FEARFUL vs. HAPPY: Evolved prosodic solutions for 2-word sentence from 29 runs each (μ values)



(a) *Int* of 3wd-sent

(b) *F0Med* of 3wd-sent

(c) *Dur* of 3wd-sent

Figure 5.20: FEARFUL vs. HAPPY: Evolved prosodic solutions for 3-word sentence from 29 runs each (μ values)

shorter *Dur* than HAPPY across the two-word sentence, but the difference was only significant for the initial word for the three-word sentence.

The differences between the two case studies may be due to FEARFUL and HAPPY sharing more prosodic space. Another possibility is that the particular prosodic variables manipulated better corresponded to distinctions between ANGRY and SAD, whereas certain other emotions may depend to a larger extent on other prosodic or acoustic cues. Also, it has been suggested before that certain emotions may be less acoustically expressed and more conveyed by other modalities. Quite often, SAD and ANGRY speech patterns are reported with more satisfactory results, whereas the scales tend to tip the other way around for for example HAPPY speech, e.g. (Burkhardt and Sendlmeier 2000). Naturally, this may merely be related to a bias in the prosodic factors examined, but on cannot ignore the impact of modality; a smile is simply a powerful cue.

Description of resynthesized mean solutions

In addition to the above statistical and graphical analyses, the utterances were resynthesized for each emotion given each prosodic gene allele's average[39] of the evolved solutions' 29 μ values for an utterance.[40] Given the experimenter's opinion, the results varied by emotion, and were overall rendered non-extreme. This could point to the importance of research on more natural emotional expression. Naturally, this may also reflect a levelling effect across participants, perhaps rooted in varying preferences. With these values, the experimenter noted some utterance differences and felt that SAD and FEARFUL appeared most distinguishable for both sentences. In addition, ANGRY was marked by increased intensity, whereas SURPRISED seemed distinguishable for the two-word sentence, but less so for the three-word sentence which may not have particularly fit this emotion semantically. HAPPY seemed least differentiable.

5.4.4 Concluding summary

To summarize, the extended study with its comparison across two sentence stimuli and across multiple words showed both reoccurring trends as well as interesting variation across words and sentences. The study demonstrated that each prosodic variable differed statistically for the factor emotion for each word examined. It also provided additional support for the positioning of ANGRY and SAD prosodic profiles relative to each other, mostly echoing the pilot study. Moreover, the results from the extended study, with an expanded set of emotions and participants, showed increased complexity. A number of interesting tendencies were

[39] Resynthesizing particular solutions, rather than based on their average, could naturally have resulted in different perceptions.

[40] These resynthesized utterances were generated on a different Linux than used in the experiment with an implementation believed faithful to the experiment.

described regarding the relative positioning and directions of the change in prosodic parameters across the utterance, some having local characteristic. The local approach also descriptively revealed local variation, both by word and by sentence, which confirmed that emotion in speech is a hard problem to address and subject to complex diversity. I believe that a larger set of possibilities and patterns may emerge if examining more utterances. Continued exploration is needed to better understand the multifaceted palette of emotional prosody. Also, whereas variation should come as no surprise (it is likely to occur in production studies and a recorded corpus as well), this study exemplified the complexity of this interesting variability, by considering spread across sentences, words, sentence locations, and prosodic parameters.[41]

There may be technical, linguistic, and prosodic motivations for the prosodic variation examined in the extended study, and the more subtle resynthesized average solutions, the effects of which seem worthy of empirical exploration in future investigations. First, the two sentences were of different types; i.e. the two-word utterance had a question-like character, whereas the three-word sentence had vocative-like features. Second, the sentences' specific words may have had an effect, not just microprosodically, but also in terms of their information-heaviness and how this interacted with prosody. For example, the three-word sentence ended vocatively with the proper name *Brownie*. Third, the sentences and their individual words had different initial *original* prosodic characteristics and contours, a limited set of fundamental prosodic parameters were examined, and Praat was used for resynthesis; the method for manipulating speech may play a role. Fourth, the fixed individual for the search process seemed important since ANGRY and SAD both showed rather stable behaviors as well as a large ratio of top-ranked fixed individuals. Fifth, the number of evaluations more than doubled for the larger 3-word problem, and more pairwise judgements may have entailed both an advantage, i.e. providing more options to the user, as well as the disadvantage of increasing evaluators' fatigue. Also, the tasks for the two-word sentence were completed before the three-word sentence and one cannot exclude that this decreased participant alertness and affected the second set of tasks. Further experimental work on impact of user type and other user variables would be interesting.

Lastly, although aiGA-theoretical questions go beyond a linguistic study, it seems valuable to mention a few thoughts in this context. One important issue is setting a run's number of evaluations. This study's decision was guided by perceived and practical limits. However, participants may differ in their tolerance limit or enter the experimental setting under different conditions.[42] Another approach for determining stop conditions, e.g. based on formal inference such as user contradiction statistics or considering physical

[41] In addition, if one would like to explore obtained parameters, one possibility would be to use a neutral TTS system to render spoken sentences, in conjunction with a copy-synthesis approach to modify them emotively (with the obvious difference that the original voice then would be synthetic, which naturally could affect the outcome). Moreover, an adjustment method would be needed to address the parameters for novel sentences, compared to those involved in the experiment. This remains a separate problem.

[42] For instance, a participant who has been awake all night, may feel quite fatigued very quickly.

signals, e.g. participant's eye-gaze or clicking rate as cues for monitoring alertness, or simply requesting fatigue estimations by the subjects may perhaps present alternative information for determining evaluation limits dynamically. In addition, desensitization is not necessarily the same phenomenon as fatigue, and it remains a research question how to establish if desensitization effects occur for a subjective judgement task, and if so, how to combat them. Lastly, it may be interesting to explore the implications of the fixed individual on aiGA's convergence patterns and perceived solution success, as well as the advantages and disadvantages it may have for restricting or guiding the search.

Chapter 6

Future work

The work presented in the previous chapters has addressed affect in both text and speech. While the presented work made contributions to the area of affective language research, it also served as a starting point for future research efforts in these still rather unexplored areas, both as regards continuing work, improvements, and extensions, as well as in opening up new research areas and applications.

First, a set of bullet points summarizes some suggestions for future work in relation to the topics more specifically discussed in this dissertation.

- Affective prosodic or other language studies often tend to advocate particular methods to emotional description (e.g. dimensions or categories) or perhaps arguing that one forms part of the other. Within a holistic approach, it instead seems interesting to further explore the mappings between different representations of emotion in speech and language, as well as what are complementary aspects gained from incorporating diverse representations into a schema of affective representation. Moreover, the question on impact from distance between emotions, e.g. in terms of dimensional properties such as quality, valence, and arousal (Banse and Scherer 1996; Scherer 2003) also deserves increased exploration.

- Given that an evaluation comparison data set for this and similar natural language processing tasks is highly subjective, not only the concept of "ground truth" may need to be reassessed, but also the evaluation measures. For subjective natural language processing tasks, it may be needed to also consider alternative evaluation techniques, such as measuring user satisfaction on applications (as Liu, Lieberman, and Selker (2003) did for an affective email client), or measuring overall system performance improvement within another task. This does not necessarily imply that evaluation techniques with controlled labeled data sets are unhelpful, but serves as a reminder that triangularizing controlled evaluation of performance together with alternative evaluation methods may provide more holistic evaluation information.

- Exploring variational properties of affective language behaviors through empirical comparison of for example sociolinguistic variables, such as gender, age, ethnicity, social class, and language background, could add additional insights.

- More exploration is needed to determine what aspects of an emotional sentence or text passage need to be synthesized with expressive voice, involving questions such as should expressive synthesis effort be concentrated on certain affect-loaded words or phrases (and, if so, how should these be characterized), or what is the relationship between direct speech and affective intensity, etc. This seems related to the exploration of locality effects on affective language cues and expressive prosody.

- If expressive storytelling is characterized as a complex bundle of several phenomena, these deserve exploration in automated text analysis as well as speakers' preferences and implementation categories and strategies. For example, automatic voice-assignment to synthetic characters (such as giving a male talking moose a dark voice but a female mouse a light one), pausing, the use of poetic devices, as well as suspense and excitement, seriousness, and so on seem to deserve attention. Also, a possible area for exploration is affective connections to Proppian story functions or other formulaic narrative structures which have been suggested by theorists.[1]

- In terms of corpus annotation in general or affect perception or production in text and speech in particular, the decision-making behavior of annotators, speakers, listeners or readers deserve future comparison, guided by the theory for design of experiments. Qualitative surveys or interviews may give additional insights on annotator behavior.

- The aiGA approach could also be used to explore other prosodic variables or narrative stances (e.g. excitement/suspense). Beyond the set of prosodic factors examined in Ch. 5, pause insertion, different F0 contours or aspects of voice quality and other prosodic effects could be considered.

- Scholars may find it interesting to quantitatively study the problem of automatic affect prediction in text with alternative classification algorithms, or rule combination (e.g. for applying binary learners as Koppel and Schler (2006) did for sentiment analysis). An important linguistic area of future improvement is to better understand the importance as well as interaction of various feature types and individual features for affect prediction in text, e.g. using controlled experiments or feature selection techniques toward this goal. Effects of alternative feature sets could also prove insightful; e.g. with text characteristics drawn from analysis of psychiatric disorders (Andreasen and Pfohl 1976; Fraser, King, Thomas, and Kendell 1986), phonoemotional profiling (Whissell 2000), or prototypical personalities or sociolinguistic variables of story characters. In addition, WordNet Affect was rather recently rere-

[1] A small trial with Potter stories indicated that story functions appeared difficult to annotate. However, this could relate to them applying more to some stories than others, or that such annotation requires specialists.

leased in an extended version; it could not be incorporated into the current text classification methods due to time-constraints, but seems like an interesting resource to explore for feature set creation for machine learning or improvement of the vocabulary method *lextag*. The *lextag* method could possibly also be improved by manual inspection and lexicographic adjustment of its word lists. Moreover, affect intensity and trajectories, as well as affect blends, decay, and interpolation phenomena in texts or speech materials, e.g. (Liu, Lieberman, and Selker 2003), represent possible directions for future work. For example, a model from the HMM family may provide insights for affective sequence and decay phenomena, as suggested by Picard (1997). In addition, empirical cross-corpora and cross-genre experimentation could reveal how affect methods trained in one area generalize across the spectrum of domains and authors, whereas analysis of error patterns could isolate pockets of interests for future improvements. Lastly, an interesting additional machine learning method could draw more specifically on continuous features, e.g. on the basis of numerical scores as features, for instance using a machine learning approach such as a CART tree which relates well with both discrete and continuous features. Given the current classification setup, any added individual methods could, on the one hand, be contrasted against other methods, and, on the other hand, be included in the combination step.

- Lastly, as noted in Ch. 3, affective semantics is difficult for most automatic techniques to capture because rather than surface features, it requires deep natural language understanding, drawing on subjective knowledge and experience. Liu, Lieberman, and Selker (2003), who voiced critique against statistical learning at the smaller sentence-level, drew on an iterative propagation over a common-sense database to approach affect inference. At the same time, approaches based on knowledge databases involve a whole set of issues such as cost in for example collection time, as well as the artificiality and limitations of trying to enumerate rather than perceiving and experiencing human knowledge. To tackle the problem of knowledge and deep understanding, affect inference via evolution or continuous online learning of artificial agents may be a promising future direction. Perhaps one alternative is using public spaces where people may choose to interact directly with agents, e.g. to avoid saturation or fatigue effects.

In addition to the above suggestions, which more specifically related to the work presented in previous chapters, this dissertation has also spawned ideas expanding the spectrum of applications and problems for text and speech research, a few of which I find particularly interesting to pursue, and which are briefly introduced below.

- Materials comparing and contrasting cross-linguistic and cross-cultural differences as well as similarities in affective language use could be beneficial if incorporated into the second language classroom. Structured incorporation of subjective language phenomena into second and foreign language learning demands further cross-cultural study of these phenomena, as well as the development of sound pedagogical materials and scholarly inquiry into appropriate teaching methods.

- Although ethics has been discussed within the field of affective computing more generally by Picard (1997), it is a relatively unexplored area in applying affective language technology. However, such guidelines are important for assuring individual rights, as affective language technologies continue to be refined as well as expanded into new domains.
- The use of robust evolutionary methods could be expanded to other linguistics problems. For instance, an especially challenging problem of interest to me is second language learning; state-of-the-art computer-assisted language learning (CALL) applications (this generally applies also to other learning materials) generally bundle language learners into a homogeneous group, when in fact learners are vastly individualized. They encounter different problem areas in language learning, apply personalized strategies to acquisition, and have a myriad of different needs, motivations, and learning goals. CALL applications based on an interactive evolutionary method such as the aiGA approach introduced in Ch. 5 could provide an innovative alternative to current methods. Within such a framework a system might evolve a learner's customized material exposure and allow a learner to interact with computational learning tools that address her learning curve in an individualized manner.
- Cross-cultural aspects of affective prosody definitely require more work. Most cross-cultural work so far has contrasted industrialized speakers. However, drawing a parallel to Ekman and colleagues' facial expression studies across Westernized and isolated populations, pursuing cross-linguistic work with remote communities seems particularly interesting. Within this context, one could explore how more isolated communities, for example the Shuar or Huarani/Woadani in the Ecuadorian Amazon jungle, perceive expression of affective prosody from other parts of the world and vice-versa.

To sum up, it is hoped that the current work will not only have contributed to the present knowledge of affect in text and speech, but that it also has served to stimulate further research on affective and subjective language. It is an exciting time for affective language computing, and the above discussion introduced a number of suggestions for facets or areas for continued or novel exploration for affective language phenomena. In the next and last chapter, I proceed to conclude the dissertation by summarizing the work and its contributions.

Chapter 7

Conclusion

This study has made an attempt at exploring the topic of affect in natural language. As mention in the introduction, the impetus for such research involves both a need for scientific exploration of affect in text and speech, as well as several useful application areas, such as improving human-computer interfaces within the area of affective computing, developing expressive text-to-speech synthesis, conducting subjective natural language processing tasks going beyond sentiment analysis, building educational dialog systems, handling language-based customer service data, exploring affective meaning dimensions for improving text summarization or machine translation, and including affective semantics for foreign language learning, just to mention a few application areas.

Set within an interest area of expressive text-to-speech synthesis, the study focused on two highly challenging problems; one related to text analysis and the other to exploring the emotional aspect of speech's prosody or voice inflection.

The first task concerned *automatic affect prediction in text*, and the research task was broadly put as: *given an input text, how and in how far is it possible to identify, i.e. predict, affect of a given sentence.* The approach taken was that of classifying sentences for affect labels, given a multi-class and multi-level scenario. The text domain was tales, keeping in mind the use of this technology for synthesized expressive storytelling (although it was also noted that affect is only one of multiple phenomena which may be involved in such oral narrative events). Specifically, several classification methods were explored, as were various approaches for merging methods output into a joint prediction, with a particular emphasis on machine learning methods. Moreover, this part of the study had interesting "side-effects". It involved the setup and discussion of a hierarchical affect model, and as comparison standard and resource for developing certain methods, a newly developed annotated corpus was described, enabling large-scale quantitative exploration of affective language phenomena. In addition, the study preferred the notion of *acceptability*, avoiding the concept of *ground truth*, since the latter seems questionable given the subjectivity of human affect perception and language phenomena. Moreover, the task and the corpus data also spawned the devising of a specific methodological machine learning setup, e.g. exploring random data set partitions for tuning and slightly adjusting k-fold cross-validation to consider the average classification performance on stories. The corpus naturally divided itself into three different subcorpora, which were explored individually, and also gave an idea of the variational performance patterns across subcorpora and affect levels. An interesting subset of sentences marked by high agreement on the affect labels received special descriptive and empirical attention, and the behavior of considering or not the knowledge of previous affect history was explored.

The second task *evolving emotional prosody with interactive evolutionary computation* concerned the speech side of language. The research task involved: *How does a set of fundamental prosodic parameters (pitch or F0, duration, and intensity) behave in emotional speech?* The method used required no costly speech database. Instead, applying a sophisticated search algorithm, where human users interactively evaluated resynthesized or computer-modified speech, speech was evolved with different emotions, and the study compared evolved behavior across emotions and against the original recordings, as well as discussed variational evidence. For this part of the study, the interesting "side effects" involved considerations for prosody or speech research, such as an emphasis on word-level encoding and the use of different utterances which allowed both a look at local prosodic patterns and other variation phenomena, as well as ideas for the future application of interactive evolutionary methods within related or similar natural language areas.

To summarize, the main contributions of this study covered four main areas:

1. Novel data set creation of a corpus of tales annotated for affect at the sentence level

2. Setup of a classification scheme for performing automatic affect prediction in text as well as application of computational classification methods to this problem

3. Use of efficient interactive evolutionary computation to evolve and explore fundamental parameters of emotional prosody in computer-modified speech based on modest resources

4. Identifying and suggesting paths for further exploration for the study of affect in language, as well as for related areas given techniques used in the study

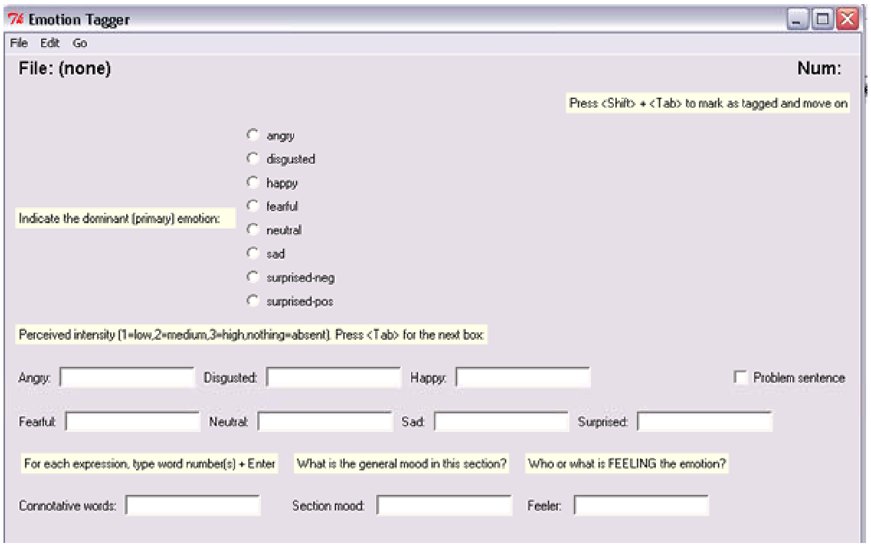
Borrowing an image from the realm of fairy tales, this study has opened up a door and taken an initial step towards exploring the world of affect in language. It will be interesting to see at which happy ending the field will eventually arrive at in its future journey. *¡Colorín, colorado, este cuento se ha acabado!*

Appendix A

Annotation manual

Manual for the emotion tagger

1. Copy the folders **dist-em-tagger0.2** and any folders containing **fairytales** from the CD to your PC, for example, onto your desktop. (**dist-em-tagger0.2** contains files and subdirectories that are needed to run the annotation program, and any **fairytales** folders contain files with the extension **.em**. These are the unannotated fairytales, and as you'll notice the names of the files correspond to the names of the fairytales.)
2. To start the emotion tagger, click on the icon in the folder called **emotions_app.exe**
3. The tagger should open up and display the following interface:



Press <Shift> + <Tab> to mark as tagged and move on

Indicate the dominant (primary) emotion:

Perceived intensity (1=low,2=medium,3=high,nothing=absent). Press <Tab> for the next box

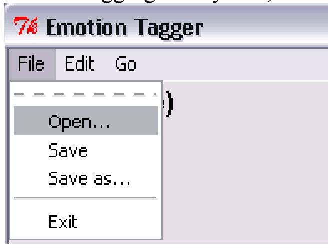
For each expression, type word number(s) + Enter

What is the general mood in this section?

Who or what is FEELING the emotion?

Note that instructions are highlighted in light-yellow on the interface.

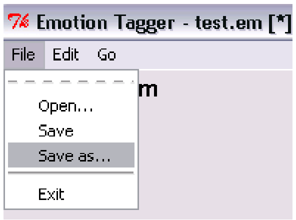
4. To start tagging a fairy-tale, click on **File** in the top menu and choose **Open**:



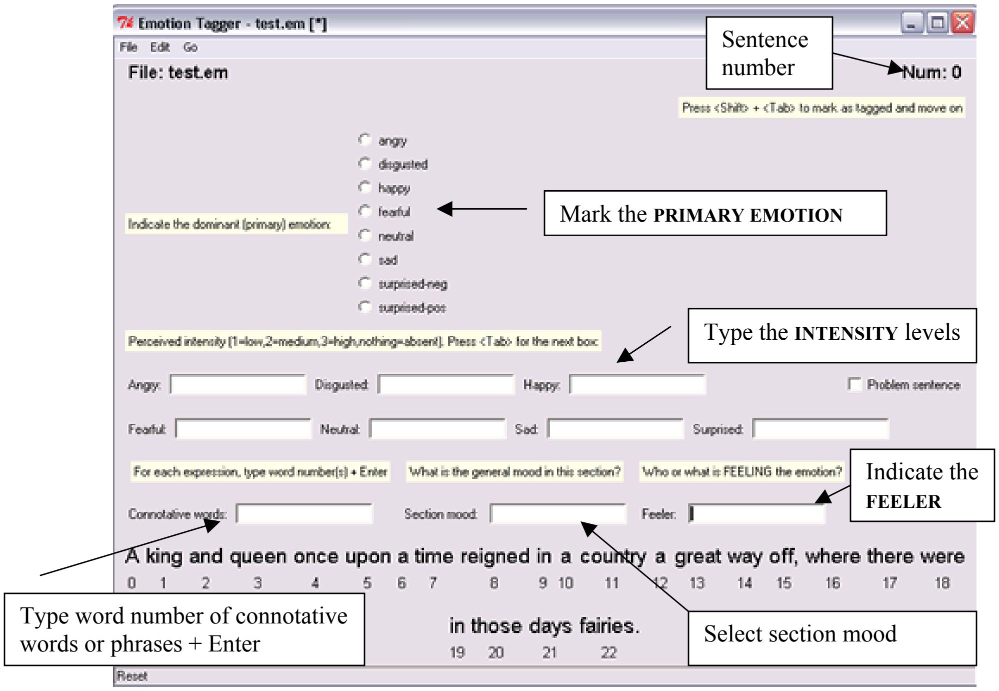
5. In the dialog box that opens next, click on a fairy-tale of your choice.

Note: A training session explained the annotation task with hands-on practice. Also, the manual was updated and corrected somewhat during the course of the annotation project (e.g. bullet 7e incorrectly said 'primary mood' in earlier version). For bullet "1", H.C. Andersen file names were abbreviated.

6. The first sentence of the fairy-tale will appear in the annotation interface. Click **File** and **Save As** to save the file under the same name but use your initials as the extension (eg: **.CA**)



7. You can begin annotating the fairy-tale's sentences. Below, each step is described.


Emotion Tagger - test.em [*]
File
Edit
Go
File: test.em
Num: 0
Press <Shift> + <Tab> to mark as tagged and move on
Indicate the dominant (primary) emotion:
○ angry
○ disgusted
○ happy
○ fearful
○ neutral
○ sad
○ surprised-neg
○ surprised-pos
Perceived intensity (1=low,2=medium,3=high,nothing=absent). Press <Tab> for the next box:
Angry:
Disgusted:
Happy:
□ Problem sentence
Fearful:
Neutral:
Sad:
Surprised:
For each expression, type word number(s) + Enter
What is the general mood in this section?
Who or what is FEELING the emotion?
Connotative words:
Section mood:
Feeler:
Reset

A king and queen once upon a time reigned in a country a great way off, where there were in those days fairies.

Sentence number

Mark the **PRIMARY EMOTION**

Type the **INTENSITY** levels

Indicate the **FEELER**

Type word number of connotative words or phrases + Enter

Select section mood

a. Choose the **PRIMARY EMOTION** in the sentence, and click on its radio button.

<u>**Important instruction:**</u>
You must be able to connect a **FEELER** with the primary emotion. Most often, this will be a character, such as the <u>speaker</u> or the character <u>performing the action</u> in the sentence. (Alterantively, you can attribute the emotion to the **READER** if the emotion is very noticable but isn't connected to a character in the story.)

b. The cursor will jump to the corresponding box for marking the **INTENSITY** of the primary emotion. Type the value that correspond to your judgement of intensity (*1=low intensity, 2= medium intensity, 3=high intensity*) .

Important instructions:
2 is the default intensity level, but you can use other intensity values to encode emotion categories that aren't included in the list of basic emotion types. For example,

- **worry** can be encoded as **low-intensity fear** (*Fear = 1*)
- **annoyed** as **low-intensity anger** (*Angry=1*), whereas
- **rage** is **high-intensity anger** (*Angry=3*), and so on.

You can use the other intensity boxes of the other emotions to mark any other secondary emotions that you feel are present in the sentence. Again, use the scale from 1 to 3. Use the **Tab**-key to jump between the boxes from now on. (You don't have to mark emotions that aren't present. These boxes will be filled with zeros automatically as you move to the next sentence. If Neutral is the primary emotion of the sentence, use a default intensity level "2" for Neutral.)

c. If you feel that any of the **words** or **phrases** in the sentence have **emotional** or **connotative meanings** that contributed to your decision making, type the number of these words. The words will appear in red font. Just verify that the correct words are in red and then press enter. The words you chose now become bold. (You can repeat this procedure several times if you want to mark other words or phrases from the sentence.)

d. Decide what's the **MOOD** of the larger section. (Use common sense and your best judgement to determine where a section of the story begins and ends.)

Important instruction:
Choose between the same categories as above for the **MOOD**, but just type the first letter of the emotion, i.e. **A**=*angry*, **D**=*disgusted*, **H**=*happy*, **F**=*fearful*, **N**=*neutral*, **S**=*sad*, **+** = *positively surprisde*, **-** = *negatively surprised*.

e. Indicate the **FEELER** of the primary emotion. Remember to prefer a character in the sentence (or in the preceeding context) as feeler, and just use **READER** in case there is no other possible attribution of the emotion.

Important instruction:
For each fairytale, you will receive a list of character names. Use these and no other terms for the characters when you indicate the **FEELER**.

f. To mark a sentence as tagged, and go on to the next sentene, just press the keys **Shift + Tab**. **NOTE: The cursor must be in the FEELER box.**

Important instruction:
You **must** type a **PRIMARY EMOTION** and a **MOOD** to go on to the next sentence. Optionally, you can flag a sentence as problematic, but minimize the use of the problem flag. We're interested in your judgements rather than "right" or "wrong".

8. When you reach the last sentence of the fairytale, you'll see a popup message. Click **Save** again.

9. To close the tagger, choose **File** and **Exit**, or just click on the **X** in the to-right corner of the application.

10. Some other features of the emotion tagger:

    a. If you want to correct a mistake, you *must* reset the sentence and retag it by choosing **Reset sentence** in the **Edit** menu. Do the following
        1. Take note of the part of the annotation which you want to keep
        2. Choose **Reset sentence** in the **Edit** menu
        3. Re-annotate the sentence.
        4. Press **Shift+Tab** with the cursor in the **FEELER** box.
        5. You can verify the new annotation (see *b* below)

    b. If you want to see a report of the current annotation of a sentence, click on **Show sentence entry** in the **Edit** menu. This will print a summary to the consol (The consol is the black screen opening up behind the tagger).

    c. You can move between sentences by using the **Go** menu and choosing **Next** or **Previous** or **Sentence Number** (for the latter, type a sentence in the box).

    d. If you open a file that has been partially tagged, you can click on **First Untagged** in the **Go** menu to jump to the first untagged sentence in the file. You will need to choose "All files" to see the file, since it now should end on your initials (instead of ".em").

    e. You can also jump to a particular sentence number by choosing **Sentence Num** (also called **Go to sentence**) in the **Go** menu.

    f. Remember, use the **Tab** key to jump between the input forms in the interface, and don't forget to mark a sentence as tagged by using **Shift + Tab.**

Contact information
If you encounter any problems or you have any questions regarding how to annotate or the annotation program, please contact Cecilia.
Email: ebbaalm@uiuc.edu .
Phone: (217) 721-7387

Again, thank you for working with us. The quality of the annotations you provide are important for the future success of this research project, and we really appreciate that you're doing your best on this decision task. If you have any feedback and comments. that you think will be useful for this research we'd be very pleased if you'd let us know.

Appendix B

Instructions in aiGA extended study

Instructions for speech perception study **User-id:** _______________

General information
This research study is conducted by Cecilia Alm (ebbaalm@uiuc.edu). It is about emotional inflection in resynthesized speech (i.e. computer-modified speech). You will use a web interface to listen to pairs of resynthesized utterances and judge which part of the pair, if any, better renders a particular target emotion in its voice inflection. Participating in this study is voluntary. After reading this document, if you want to consent to participating, please sign and date the attached form, and hand it to the experimenter before the experimental tasks start.

Range of experiment
You will participate in the experiment 1 day in the NCSA building on campus, and do several evaluation tasks. The experiment takes around 3 hours. You will be paid $20.

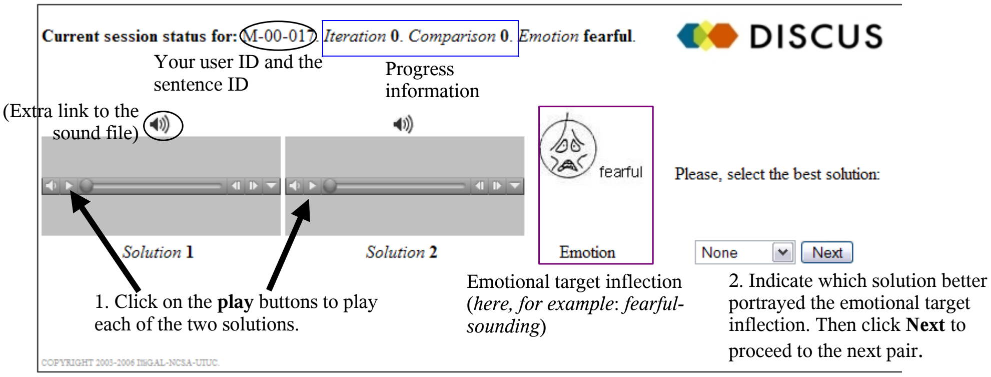
Setting yourself up for the experiment
So that the experimental conditions are always the same, please use the same laptop, do not modify the volume level, use headphones, and focus on the task. We will take a short break between each task to help you focus on a particular emotional target.

The initial web interface (first screen)

- *User ID* **should say what is written at the top of this page** (if it says *null*, or something else than your user-id, please report to the experimenter).
- *Sentence ID:* The task's sentence number should occur in the field *Sentence ID.*
- *Emotion:* Each task's *emotional target* should occur in the field *Emotion.*

Pair-wise judgment task (main screen)
For each task, you will listen to several pairs of resynthesized sentences. Each time, you are presented with a pair of sound files, called "solutions", which you should compare. You'll listen to a real sentence, but you should <u>disregard word meaning</u> and **focus on the voice and what it sounds like.**

- First, listen to the two solutions
- Next, using the drop-down menu, **indicate which solution you feel better portrays the emotional target inflection (e.g. *fearful-sounding*)**
- Lastly, click on the button labeled "**Next**" to go on to the next pair
- Each task **STOPS** after a certain number of pairs. A message will tell you when to stop.



Current session status for: M-00-017. *Iteration* **0**. *Comparison* **0**. *Emotion* **fearful**.

Please, select the best solution:

Your user ID and the sentence ID

Progress information

(Extra link to the sound file)

1. Click on the **play** buttons to play each of the two solutions.

Emotional target inflection (*here, for example*: *fearful-sounding*)

2. Indicate which solution better portrayed the emotional target inflection. Then click **Next** to proceed to the next pair.

Figure: Main screen for pair-wise comparative judgments

Instructions for speech perception study **User-id:** _______________

A few things to note:

- Please focus on the *emotional target* for each task task. You should indicate which sound file of the pair better rendered the sentence for that emotion target.
- An emotion category can take various shades and guises. For example, *fear* includes *worry*, *anxiety*, or *panic,* and they are all regarded acceptable as portraying the emotion.
- You can listen to the sound files as many times as you want to.
- Resynthesized speech may have a certain "buzziness". Please ignore any buzziness when making your judgments.
- Make sure you have selected which alternative best portrayed the emotional target inflection in the drop-down menu before clicking "*Next*".
- Choose "***None***" when you think the two sound files are **identical**, or when you think they are **equally bad.** (Don't overuse *None*, though, since what we are interested in knowing is your discriminative judgments).
- Do not click too fast, since the system needs some time to prepare the next pair. You'll always hear the same sentence in a task. If you don't hear the sound file or the whole sentence when you click the *play* button, please wait a little bit and then click *play* again. (If the problem doesn't go away, notify the experimenter.)
- Please do not return to a previous judgment (i. e. do not use the browser button to go back and change your mind).
- Each task consists of making judgments for 3 iterations (labeled 0 to 2). Each iteration has several pair-wise comparative judgments. The task stops after iteration 3, comparison 0, when you get a message that you've finished the task.
- After you finish one task, please wait at your computer station until the experimenter announces a break to not disturb the other participants. There will be short breaks between the different emotional conditions.
- After the last task, please complete the short paper survey handed to you. Before leaving the lab, please hand the paper survey to the experimenter, sign the payment form, and receive your $20 payment for participating.
- If you want a copy of this document, or would like more information about the study, please contact the experimenter. The contact information is listed below.

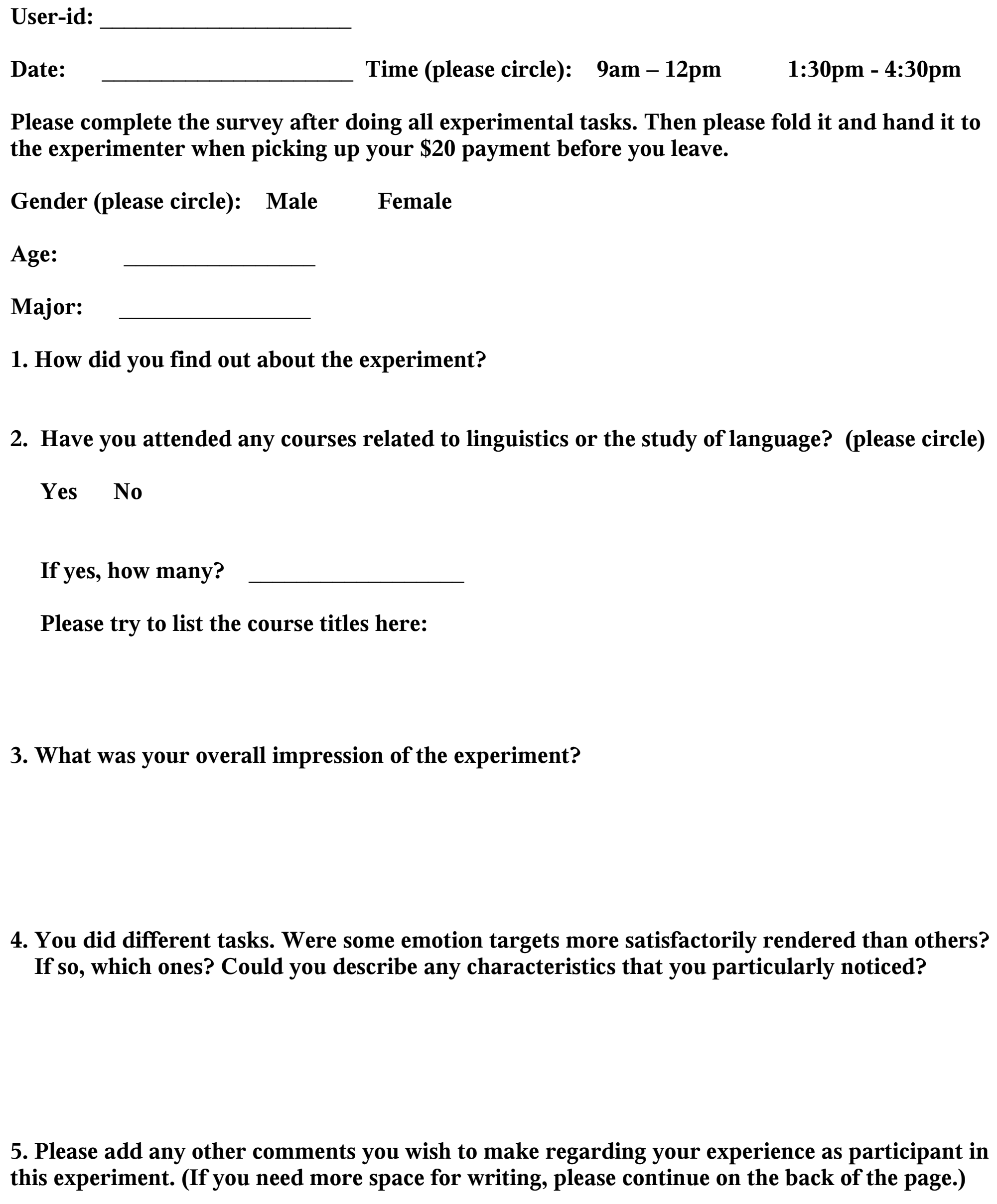
THANK YOU VERY MUCH!

Contact information:
Cecilia Ovesdotter Alm, ebbaalm@uiuc.edu, Phone: 217 - 721 7387 **IRB:** #05136
Faculty sponsors: Richard Sproat (rws@uiuc.edu), Chilin Shih (cls@uiuc.edu), Xavier Llora (xllora@uiuc.edu).

2

Appendix C

Survey in aiGA extended study



User-id: ____________________

Date: ____________________ Time (please circle): 9am – 12pm 1:30pm - 4:30pm

Please complete the survey after doing all experimental tasks. Then please fold it and hand it to the experimenter when picking up your $20 payment before you leave.

Gender (please circle): Male Female

Age: ______________

Major: ______________

1. How did you find out about the experiment?

2. Have you attended any courses related to linguistics or the study of language? (please circle)

Yes No

If yes, how many? ________________

Please try to list the course titles here:

3. What was your overall impression of the experiment?

4. You did different tasks. Were some emotion targets more satisfactorily rendered than others? If so, which ones? Could you describe any characteristics that you particularly noticed?

5. Please add any other comments you wish to make regarding your experience as participant in this experiment. (If you need more space for writing, please continue on the back of the page.)

References

Abelin, Å. and J. Allwood (2003). Cross linguistic interpretation of emotional prosody. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 110–113.

Aharonson, V. and N. Amir (2006). Emotion elicitation in a computerized gambling game. In *Proceedings of Speech Prosody.*

Alexander, K. J., P. J. Miller, and J. A. Hengst (2001). Young children's emotional attachments to stories. *Social Development 10*(3), 374–398.

Alías, F., X. Llorà, I. Iriondo, and L. Formiga (2003). Ajuste subjetivo de pesos para selección de unidades a través de algoritmos genéticos interactivos. *Procesamiento del Lenguaje Natural 31*, 75–82.

Alm, C. O. and X. Llorà (2006). Evolving emotional prosody. In *Proceedings of Interspeech*, pp. 1826–1829.

Alm, C. O., N. Loeff, and D. Forsyth (2006). Challenges for annotating images for sense disambiguation. In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora*, pp. 1–4.

Alm, C. O., D. Roth, and R. Sproat (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of HLT/EMNLP*, pp. 579–586.

Alm, C. O. and R. Sproat (2005a). Emotional sequencing and development in fairy tales. In *Proceedings of the First International Conference on Affective Computing and Intelligent Interaction*, pp. 668–674.

Alm, C. O. and R. Sproat (2005b). Perceptions of emotions in expressive storytelling. In *Proceedings of Interspeech*, pp. 533–536.

Anderson, C. W. and G. E. McMaster (1982). Computer-assisted modeling of affective tone in written documents. *Computers and the Humanities 16*, 1–9.

Andreasen, N. G. and B. Pfohl (1976). Linguistic analysis of speech in affective disorders. *Archives of General Psychiatry 33*(11), 1361–1367.

Banse, R. and K. R. Scherer (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology 70*(3), 614–636.

Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Nöth (2000). Desperately seeking emotions or: Actors, wizards, and human beings. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 195–200.

Bayerl, P. S. and K. I. Paul (2007). Identifying sources of disagreement: generalizability theory in manual annotation studies. *Journal of Computational Linguistics 33*(1), 3–8.

Bernsen, N. O., M. Charfuelàn, A. Corradini, L. Dybkjær, T. Hansen, S. Kiilerich, M. Kolodnytsky, D. Kupkin, and M. Mehta (2004). First prototype of a conversational H.C. Andersen. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, pp. 458–461.

Besnier, N. (1990). Language and affect. *Annual Review of Anthropology 19*, 419–451.

Blocher, K. and R. W. Picard (2002). Affective Social Quest: Emotion recognition theraphy for autistic children. In K. Dautenhahn et al (Ed.), *Socially Intelligent Agents - creating relationships with computers and robots*, pp. 133–140. Dordrecht: Kluwer Academic Publishers.

Blom, K. J. and S. Beckhaus (2005). Emotional storytelling. In *IEEE Virtual Reality Conference: Workshop on Virtuality Structure*, pp. 23–27.

Boersma, P. and D. Weenink (2005). Praat: doing phonetics by computer (Version 4, subversion unknown) [Computer program]. Retrieved from http://www.praat.org/.

Boersma, P. and D. Weenink (2006). Praat: doing phonetics by computer (probably Version 4.5.01) [Computer program]. Retrieved Oct. 2006, from http://www.praat.org/.

Boucouvalas, A. C. (2002). Real time text-to-emotion engine for expressive Internet communication. In G. Riva, F. Davide, and W. IJsselsteijn (Eds.), *Emerging Communication: Studies on New Technologies and Practices in Communication*, Volume 5, pp. 305–318. IOS Press: Amsterdam, The Netherlands.

Bradley, M. and P. Lang (1999). Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical Report Technical Report C-1, University of Florida, Gainesville.

Bralé, V., V. Maffiolo, I. Kanellos, and T. Moudenc (2005). Towards an expressive typology in storytelling: A perceptive approach. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 858–865.

Bühler, K. (1934). *Sprachtheorie: Die Darstellungsfunktion der Sprache.* Stuttgart: Gustav Fischer Verlag.

Burkhardt, F. and W. F. Sendlmeier (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 151–156.

Cabral, J., L. Oliveira, G. Raimundo, and A. Paiva (2006). What voice do we expect from a synthetic character? In *SPECOM*, pp. 536–539.

Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society 8*, 1–19.

Calev, A., D. Nigal, and S. Chazan (1989). Retrieval from semantic memory using meaningful and meaningless constructs by depressed, stable bipolar and manic patients. *British Journal of Clinical Psychology 28*, 67–73.

Campbell, N. (2004). Perception of affect in speech - towards an automatic processing of paralinguistic information in spoken conversation. In *Proceedings of Interspeech-ICSLP*, pp. 881–884.

Carlson, A., C. Cumby, N. Rizzolo, J. Rosen, and D. Roth (1999). SNoW user manual. Technical report, UIUC Comp. Science.

Cauldwell, R. T. (2000). Where did the anger go? The role of context in interpreting emotion in speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 127–131.

Chambers, N., J. Tetreault, and J. Allen (2004). Approaches for automatically tagging affect. In Y. Qu, J. Shanahan, and J. Wiebe (Eds.), *Exploring Attitude and Affect in Text: Theories and Applications: Papers from the 2004 Spring Symposium*, pp. 36–43. Technical Report SS-04-07. American Association for Artificial Intelligence, Menlo Park, California.

Chuenwattanapranithi, S., Y. Xu, B. Thipakorn, and S. Maneewongvatana (2006). Expressing anger and joy with the size code. In *Proceedings of Speech Prosody.*

Collier, G. (1985). *Emotional Expression.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cornelius, R. R. (2000). Theoretical approaches to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 3–10.

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

Darwin, C. (1998 [1890]). The expression of the emotions in man and animals [selected excerpts]. In J. M. Jenkins, K. Oatley, and N. L. Stein (Eds.), *Human Emotions: A Reader*, pp. 13–20. Malden, Massachussetts: Blackwell.

Dave, K., S. Lawrence, and D. M. Pennock (2003). Mining the peanut gallery: Opinion extraction and semantic classificaton of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, pp. 519–528.

Dellaert, F., T. Polzin, and A. Waibel (1996). Recognizing emotion in speech. In *Proceedings of the ICSLP*, pp. 1970–1973.

Devillers, L., S. Abrilian, and J.-C. Martin (2005). Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 519–526.

Devillers, L., I. Vasilescu, and C. Mathon (2003). Acoustic cues for perceptual emotion detection in task-oriented human-human corpus. In *15th International Congress of Phonetic Sciences*, pp. 1505–1508.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers.

Dyer, M. G. (1987). Emotions and their computations: Three computer models. *Cognition and Emotion 1*(3), 323–347.

Eide, E., A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli (2004). A corpus-based approach to <ahem/> expressive speech synthesis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 79–84.

Ekman, P. (1994). All emotions are basic. In P. Ekman and R. J. Davidson (Eds.), *The Nature of Emotion: Fundamental Questions*, pp. 15–19. Oxford: Oxford University Press.

Ekman, P. and W. V. Friesen (1998 [1971]). Constants across culture in the face and emotion. In J. M. Jenkins, K. Oatley, and N. L. Stein (Eds.), *Human Emotions: A Reader*, pp. 63–72. Malden, Massachussetts: Blackwell.

Esuli, A. and F. Sebastiani (2006). SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*, Genova, IT, pp. 417–422.

Fagel, S. (2006). Emotional McGurk effect. In *Proceedings of Speech Prosody*.

Foolen, A. (1997). The expressive function of language: Towards a cognitive semantic approach. In S. Niemeier and R. Dirven (Eds.), *The Language of Emotions: Conceptualization, Expression, and Theoretical Foundation*, pp. 15–31. Amsterdam: John Benjamins.

Francisco, V. and P. Gervás. Análisis de dependencias para la marcación de cuentos con emociones. *Procesamineto del Lenguaje Natural*. [To appear].

Francisco, V. and P. Gervás (2006a). Automated mark up of affective information in English texts. In *TSD*. [To appear].

Francisco, V. and P. Gervás (2006b). Exploring the compositionality of emotions in text: Word emotions, sentence emotions and automated tagging. In *AAAI-06 Workshop on Computational Aesthetics: Artificial Intelligence Approaches to Beauty and Happiness*. [To appear].

Francisco, V., P. Gervás, and R. Hérvas (2005). Análisis y síntesis de expresión emocional en cuentos leídos en voz alta. *Procesamineto del Lenguaje Natural 35*, 293–300.

Fraser, W. I., K. M. King, P. Thomas, and R. E. Kendell (1986). The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry 148*, 275–278.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *COLING*, pp. 841–847.

Généreux, M. and R. Evans (2006). Distinguishing affective states in weblog posts. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 40–42.

Gobl, C. and A. Ní Chasaide (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication 40*(1-2), 189–212.

Goddard, C. (1998). *Semantic Analysis: A Practical Introduction*. Oxford: Oxford University Press.

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, Massachusetts: Addison Wesley.

Goldberg, D. E. (2002). *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Boston: Kluwer Academic Publishers.

Gonzáles, C., J. A. Lozano, and P. Larrañaga (2002). Mathematical modelling of UMDA$_c$ algorithm with tournament selection. Behaviour on linear and quadratic functions. *International Journal of Approximate Reasoning 31*(3), 313–340.

Gordon, A., A. Kazemzadeh, A. Nair, and M. Petrova (2003). Recognizing expressions of commonsense psychology in English text. In *Proceedings of ACL*, pp. 208–215.

Gustafson, J. and K. Sjölander (2003). Voice creation for conversational fairy-tale characters. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 145–150.

Halliday, M. A. K. (1996). Linguistic function and literary style: An inquiry into the language of William Golding's `The Inheritors`. In J. J. Weber (Ed.), *The Stylistics Reader: From Roman Jakobson to the Present*, pp. 56–86. London: Arnold.

Harvey, P. D. (1983). Speech competence in manic and schizophrenic psychoses: The association between clinically rated thought disorder and cohesion and reference performance. *Journal of Abnormal Psychology 92*(3), 368–377.

Hasegawa-Johnson, M., S. E. Levinson, and T. Zhang (2004). Children's emotion recognition in an intelligent tutoring scenario. In *Proceedings of Interspeech*, pp. 1441–1444.

Hatzivassiloglou, V. and K. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of ACL*, pp. 174–181.

Hedquist, R. (1978). *Emotivt språk: En studie i dagstidningarnas ledare*. Ph. D. thesis, Umeå.

Hoey, M. (2000). Persuasive rhetoric in linguistics: A stylistic study of some features of the language of Noam Chomsky. In S. Hunston and G. Thompson (Eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse*, pp. 28–37. Oxford: Oxford University Press.

Hofer, G. O., K. Richmond, and R. A. J. Clark (2005). Informed blending of databases for emotional speech synthesis. In *Proceedings of Interspeech*, pp. 501–504.

Hoffman, R. E., S. Stopek, and N. C. Andreasen (1986). A comparative study of manic vs schizophrenic speech disorganization. *Archives of General Psychiatry 43*(9), 831–838.

Holzman, L. E. and W. Pottenger (2003). Classification of emotions in Internet chat: An application of machine learning using speech phonemes. Technical Report LU-CSE-03-002, Leigh University.

Huang, X., Y. Yang, and C. Zhou (2005). Emotional metaphors for emotion recognition in Chinese text. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 319–325.

Iriondo, I., F. Alías, J. Melenchón, and M. A. Llorca (2004). Modeling and synthesizing emotional speech for Catalan text-to-speech synthesis. In *Affective Dialogue Systems*, pp. 197–208.

Jakobson, R. (1996). Closing statement: Linguistics and poetics. In J. J. Weber (Ed.), *The Stylistics Reader: From Roman Jakobson to the Present*, pp. 10–35. London: Arnold.

John, D., A. C. Boucouvalas, and Z. Xu (2006). Representing emotional momentum within expressive internet communication. In *Proceedings of the 24th IASTED International Conference on Internet and Multimedia Systems and Applications*, pp. 183–188.

Johnson-Laird, P. N. and K. Oatley (1989). The language of emotions: an analysis of a semantic field. *Cognition and Emotion 3*(2), 81–123.

Karla, A. and K. Karahalios (2005). TextTone: Expressing emotion through text. In *Interact 2005*, pp. 966–969.

Ke, J., M. Ogura, and W. S.-Y. Wang (2003). Optimization models of sound systems using Genetic Algorithms. *Journal of Computational Linguistics 29*(1), 1–18.

Keltner, D. and P. Ekman (2003). Introduction to expression of emotion. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*, pp. 411–414. Oxford: Oxford University Press.

Kim, J., E. André, M. Rehm, T. Vogt, and J. Wagner (2005). Integrating information from speech and physiological signals to achieve emotional sensitivity. In *Proceedings of Interspeech*, pp. 809–812.

Klabbers, E. and J. P. H. van Santen (2004). Clustering of foot-based pitch contours in expressive speech. In *Proceedings of the 5th ISCA Speech Synthesis Workshop*, pp. 73–78.

Kochanski, G. and C. Shih (2003). Prosody modeling with soft templates. *Speech Communication 39*(3-4), 311–352.

Koomen, P., V. Punyakanok, D. Roth, and W.-T. Yih (2005). Generalized inference with multiple semantic role labeling system. In *Proceedings of the Annual Conference on Computational Language Learning*, pp. 181–184.

Koppel, M. and J. Schler (2006). The importance of neutral examples for learning sentiment. *Computational Intelligence Journal Special Issue on Sentiment Analysis 22*(2), 100–109.

Labov, W. (1973). The boundaries of words and their meaning. In C.-J. Bailey and R. W. Shuy (Eds.), *New Ways of Analyzing Variation in English*, pp. 340–373. Washington, D.C.: Georgetown Univ. Press.

Lee, L. (2004). A matter of opinion: Sentiment analysis and business intelligence (position paper). In *IBM Faculty Summit on the Architecture of On-Demand Business*.

Lee, S., S. Yildirim, A. Kazemzadeh, and S. Narayanan (2005). An articulatory study of emotional speech production. In *Proceedings on Interspeech*, pp. 497–500.

Linnankoski, I., L. Leinonen, M. Vihla, M.-L. Laakso, and S. Carlson (2005). Conveyance of emotional connocations by a single word in English. *Speech Communication 45*, 27–39.

Lipton, E. (2006, October 4). Software being developed to monitor opinions of U.S. The New York Times. Newspaper article.

Liscombe, J., G. Richard, and D. Hakkani-Tür (2005). Using context to improve emotion detection in spoken dialog systems. In *Proceedings of Interspeech*, pp. 1845–1848.

Litman, D. and K. Forbes-Riley (2004a). Annotating student emotional states in spoken tutoring dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*.

Litman, D. and K. Forbes-Riley (2004b). Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of ACL*, pp. 351–358.

Liu, H., H. Lieberman, and T. Selker (2003). A model of textual affect sensing using real-world knowledge. In *International Conference on Intelligent User Interfaces*, pp. 125–132.

Liu, H. and P. Singh (2004). ConceptNet - a practical commonsense reasoning tool-kit. *BT Technology Journal 22*(4), 211–226.

Llorà, X., F. Alías, L. Formiga, K. Sastry, and D. E. Goldberg (2005). Evaluation consistency in iGAs: User contradictions as cycles in partial-ordering graphs. Technical Report IlliGAL TR No 2005022, UIUC.

Llorà, X., K. Sastry, D. E. Goldberg, A. Gupta, and L. Lakshmi (2005). Combating user fatigue in iGAs: Partial ordering, Support Vector Machines, and synthetic fitness. Technical Report IlliGAL TR No 2005009, UIUC.

Loeff, N., C. O. Alm, and D. Forsyth (2006). Discriminating image senses by clustering with multimodal features. In *Proceedings of ACL*, pp. 547–554.

Loveland, K. A. (2005). Social-emotional imparment and self-regulation in Autism spectrum disorders. In J. Nadel and D. Muir (Eds.), *Typical and Impaired Emotional Development*, pp. 365–382. Oxford: Oxford University Press.

Lyons, J. (1977). *Semantics*, Volume 1, 2. Cambridge: Cambridge University Press.

Ma, C., H. Prendinger, and M. Ishizuka (2005). Emotion estimation and reasoning based on affective textual interaction. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 622–628.

Mathieu, Y. Y. (2005). Annotation of emotions and feelings in text. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 350–357.

Mihalcea, R. and H. Liu (2006). A corpus-based approach to finding happiness. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Min Lee, C. M., S. S. Narayanan, and R. Pieraccini (2002). Combining acoustic and language information for emotion recognition. In *Proc. of ICSLP*.

Murray, I. R. and J. L. Arnott (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America 93*(2), 1097–1108.

Murray, I. R. and J. L. Arnott (1996). Synthesizing emotions in speech: Is it time to get excited? In *Proceedings of ICSLP*, pp. 1816–1819.

Murray, I. R., M. D. Edginton, D. Campion, and J. Lynn (2000). Rule-based emotion synthesis using concatenated speech. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pp. 173–177.

Nass, C., U. Foehr, S. Brave, and M. Somoza (2001). The effects of emotion of voice in synthesized and recorded speech. In *Proceedings of the AAAI Symposium Emotional and Intelligent II: The Tangled Knot of Social Cognition*.

Niemeier, S. and R. Dirven (1997). *The Language of Emotions: Conceptualization, Expression, and Theoretical Foundation*. Amsterdam: John Benjamins.

Ogilvie, D. M., P. J. Stone, and E. S. Shnediman (1969). Some characteristics of genuine versus simulated suicide notes. *Bulletin of Suicidology*, 27–32.

Ortony, A., G. Clore, and A. Collins (1988). *The Cognitive Structure of Emotions*. Cambridge: Cambridge University Press.

Osgood, C. E. (1969). On the whys and wherefores of E, P, and A. *Journal of Personality and Social Psychology 12*(3), 194–199.

O'Shaughnessy, D. (2000). *Speech Communications: Human and Machine* (2nd ed.). New York: IEEE Press.

Oudeyer, P.-Y. (2003). The production and recognition of emotions in speech: Features and algorithms. *International Journal of Human-Computer Studies 59*(1-2), 157–183.

Owsley, S., S. Sood, and K. J. Hammond (2006). Domain specific affective classification of documents. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Palmer, G. B. and D. J. Occhi (1999). *Languages of Sentiment: Cultural Constructions of Emotional Substrates*. Amsterdam: John Benjamins.

Pang, B. and L. Lee (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*, pp. 271–278.

Pang, B. and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pp. 115–124.

Pang, B., L. Lee, and S. Vaithyanathan (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pp. 79–86.

Peinado, F., P. Gervás, and B. Díaz-Agudo (2004). A description logic ontology for fairy tale generation. In *Proceedings of the Workshop on Language Resources for Linguistic Creativity, LREC*, pp. 56–61.

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

Polzin, T. S. and A. H. Waibel (1998). Detecting emotions in speech. In *Proceedings of the CMC*.

Precht, K. (2000, August). *Patterns of Stance in English*. Ph. D. thesis, Northern Arizona University.

Propp, V. A. (1968). *Morphology of the Folktale* (2nd ed.). Austin: Univ. Texas Press.

Ragin, A. B. and T. F. Oltmanns (1983). Predictability as an index of imparied verbal communication in schizophrenic and affective disorders. *British Journal of Psychiatry 143*, 578–583.

Ragin, A. B. and T. F. Oltmanns (1987). Communication and thought disorder in schizophrenics and other diagnostic groups. a follow up study. *British Journal of Psychiatry 150*, 494–500.

Read, J. (2005). Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop*, pp. 43–48.

Redford, M. A., C. C. Chen, and R. Miikkulainen (1998). Modeling the emergence of syllable systems. In *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*, pp. 882–886.

Reilly, J. and L. Seibert (2003). Language and emotion. In R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Eds.), *Handbook of Affective Sciences*, pp. 535–559. Oxford: Oxford University Press.

Ries, K. and A. Waibel (2001). Activity detection for information access to oral communication. In *HLT*, pp. 1–6.

Rotaru, M. and D. Litman (2005). Using word-level pitch features to better predict student emotions during spoken tutoring dialogues. In *Proceedings of Interspeech*, pp. 881–884.

Roy, D. and A. Pentland (1996). Automatic spoken affect analysis and classification. In *Proceedings of the 2nd International Conference on Automatic Face and Gesture Recognition*, pp. 363–367.

Russell, J. A. and J. M. Fernández-Dols (1998 [1997]). What does a facial expression mean? [selection]. In J. M. Jenkins, K. Oatley, and N. L. Stein (Eds.), *Human Emotions: A Reader*, pp. 73–77. Malden, Massachussetts: Blackwell.

Sato, Y. (2005). Voice quality conversion using interactive evolution of prosodic control. *Applied Soft Computing 5*, 181–192.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication 40*(1-2), 227–256.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Proceedings of EUROSPEECH*, pp. 561–564.

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication 40*(1-2), 99–116.

Schröder, M. (2004). Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In *Affective Dialogue Systems*, pp. 209–220.

Schuller, B., R. Müller, M. Lang, and G. Rigoll (2005). Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In *Proceedings of Interspeech*, pp. 805–808.

Shih, C. *Prosody Learning and Generation*. Berlin: Springer. [Forthcoming].

Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Proceedings of Interspeech*, pp. 1781–1784.

Silva, A., G. Raimundo, and A. Paiva (2003). Tell me that bit again... bringing interactivity to a virtual storyteller. In *International Conference on Virtual Storytelling*, pp. 146–154.

Sproat, R. (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Dordrecht: Kluwer Academic Publishers.

Strappavara, C. and A. Valiutti (2004). WordNet-Affect: An affective extension of WordNet. In *LREC*, pp. 1083–1086.

Strappavara, C., A. Valiutti, and O. Stock (2006). The affective weight of lexicon. In *LREC*.

Subasic, P. and A. Huettner (2000). Affect analysis of text using fuzzy semantic typing. In *The 9th IEEE International Conference on Fuzzy Systems*, pp. 483–496.

Taboada, M. and J. Grieve (2004). Analyzing appraisal automatically. In Y. Qu, J. Shanahan, and J. Wiebe (Eds.), *Exploring Attitude and Affect in Text: Theories and Applications: Papers from the 2004 Spring Symposium*, pp. 158–161. Technical Report SS-04-07. American Association for Artificial Intelligence, Menlo Park, California.

Takagi, H. (2001). Interactive evolutionary computation: Fusion of the capabilities of EC optimization and human evaluation. *Proceedings of the IEEE 89*(9), 1275–1296.

Taylor, J. R. (1995). *Linguistic Categorization: Prototypes in Linguistic Theory* (2nd ed.). Oxford: Oxford University Press.

Theune, M., S. Faas, A. Nijholt, and D. Heylen (2004). The virtual storyteller: Story creation by intelligent agents. In *Proceedings Technologies for Interactive Digital Storytelling and Entertainment*, pp. 204–215.

Theune, M., K. Meijs, D. Heylen, and R. Ordelman (2006). Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech, and Language Processing 14*(4), 1137–1144.

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

van Santen, J. P. H., L. Black, G. Cohen, A. B. Kain, E. Klabbers, T. Mishra, J. de Villiers, and X. Niu (2003). Applications of computer generated expressive speech for communication disorders. In *Proceedings of Eurospeech*, pp. 1657–1660.

Vidrascu, L. and L. Devillers (2005). Real-life emotion representation and detection in call centers data. In J. Tao, T. Tan, and R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction, First International Conference, ACII 2005, Beijing, China, October 22-24, 2005, Proceedings*, pp. 739–746.

Wennerstrom, A. (2001). *The Music of Everyday Speech: Prosody and Discourse Analysis.* Oxford: Oxford University Press.

Whissell, C. (1989). The dictionary of affect in language. In R. Plutchik and H. Kellerman (Eds.), *Emotion: Theory, Research and Experience, vol 4: The Measurement of Emotions*, pp. 113–131. New York: Academic Press.

Whissell, C. (1999). Phonosymbolism and the emotional nature of sounds: Evidence of the preferential use of particular phonemes in texts of differing emotional tone. *Perceptual and Motor Skills 89*(1), 19–48.

Whissell, C. (2000). Phonoemotional profiling: A description of the emotional flavour of English texts on the basis of the phonemes employed in them. *Perceptual and Motor Skills 91*(2), 617–648.

Whissell, C. (2001). Sound and emotion in given names. *Names 49*(2), 97–120.

Whitelaw, C., N. Garg, and S. Argamon (2005). Using appraisal taxonomies for sentiment analysis. In *Midwestern Conference on Computational Linguistics.*

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

Wierzbicka, A. (1986). Human emotions: Universal or culture-specific? *American Anthropolist 88*(3), 584–594.

Wilson, T. and J. Wiebe (2003). Annotating opinions in the world press. In *4th SigDial workshop on Discourse and Dialogue.*

Wilson, T., J. Wiebe, and P. Hoffman (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of HLT/EMNLP*, pp. 347–354.

Wilson, T., J. Wiebe, and R. Hwa (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pp. 761–769.

Winton, W. M. (1990). Language and emotion. In H. Giles and W. P. Robinson (Eds.), *Handbook of Language and Social Psychology*, pp. 33–49. Chichester: John Wiley & Sons Ltd.

Yu, H. and V. Hatzivassiloglou (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of EMNLP*, pp. 129–136.

Zaenen, A. (2006). Mark-up barking up the wrong tree. *Journal of Computational Linguistics 32*(4), 577–580.

Zhang, J. Y., A. W. Black, and R. Sproat (2003). Identifying speakers in children's stories for speech synthesis. In *Proceedings of Eurospeech*, pp. 2041–2044.

Zhang, T., M. Hasegawa-Johnson, and S. E. Levinson (2003). Mental state detection of dialogue system users via spoken language. In *ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition.*

Author's Biography

Cecilia Ovesdotter Alm was born and grew up in Sweden. As the daughter of two dedicated researchers, she and her sister quickly learned to be resourceful, for example in terms of learning how to cook at an early age when their parents were absent-mindedly working on their research.

She moved to Austria after completing high school, and studied at the University of Vienna, with some time in Spain and Pittsburgh, before attending University of Illinois at Urbana-Champaign; first on an exchange award, and then as a graduate student. After earning her M.A. in Linguistics at UIUC, she spent time in Ecuador, where she had the opportunity to serve as an instructor at Universidad San Francisco de Quito. In 2003, she returned to UIUC for continuing with the Ph.D. program. She worked for three years as a Research Assistant for Prof. Richard Sproat on topics which culminated in a dissertation about affect in text and speech modalities, relevant for applications such as text-to-speech synthesis. In addition to receiving the Ph.D. degree in Linguistics, Cecilia leaves UIUC holding the Certificate in Advanced Study of Speech and Natural Language Processing, as well as the Computing in the Sciences and Engineering Option. At UIUC, she has additionally taught linguistics and language courses and worked in web development and management.

Cecilia has a strong passion for language. Besides her native Swedish, she is fluent in Spanish, German, and English and has also studied other languages. Her interests include social and affective language variation, multimodal semantics, linguistic study of literary genres, instructional technology, foreign language teaching, and less commonly taught languages such as Scandinavian languages and the language context of Ecuador. Experimental and computational methodology sometimes shape her research, as do analytical methods, and she enjoys fieldwork. She has a vision for the Humanities becoming more linked to and interacting more with the Sciences in mutual benefit.