4th International Workshop on Corpora for Research on EMOTION SENTIMENT & SOCIAL SIGNALS ES³ 2012

Workshop Programme

09:00 – 09:10
Laurence Devillers
Opening

ORAL SESSION 1: MULTILINGUAL SENTIMENT RESOURCES AND ANALYSIS
(Chair: Paolo Rosso)

09:10 – 09:30
Andrea Gianti, Cristina Bosco, Viviana Patti, Andrea Bolioli and Luigi Di Caro
Annotating Irony in a Novel Italian Corpus for Sentiment Analysis

09:30 – 09:50
Sandi Pohorec, Ines Ceh, Peter Kokol and Milan Zorman
Sentiment Analysis Resources for Slovene Language

09:50 – 10:10
Ozan Cakmak, Abe Kazemzadeh, Dogan Can, Serdar Yildirim and Shrikanth Narayanan
Root-Word Analysis of Turkish Emotional Language

10:10 – 10:30
Minlie Huang, Lei Fang and Xiaoyan Zhu
A Chinese Corpus for Sentiment Analysis

10:30 - 11:00
COFFEE BREAK

ORAL SESSION 2: LAUGHTER AND SOCIAL SIGNAL PROCESSING
(Chair: Catherine Pelachaud)

11:00 – 11:20
Khiet Truong and Jürgen Trouvain
Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis

11:20 – 11:40
Radoslaw Niewiadomski, Jérôme Urbain, Catherine Pelachaud and Thierry Dutoit
Finding out the Audio and Visual Features that Influence the Perception of Laughter Intensity and Differ in Inhalation and Exhalation Phases

11:40 – 12:00
Gary McKeown, Roddy Cowie, Will Curran, Willibald Ruch and Ellen Douglas-Cowie
ILHAIRE Laughter Database

12:00 – 12:20
Jürgen Trouvain and Khiet Truong
Comparing Non-Verbal Vocalisations in Conversational Speech Corpora

12:20 – 12:40
Magalie Ochs, Paul Brunet, Gary McKeown, Catherine Pelachaud and Roddy Cowie
Smiling Virtual Characters Corpora

12:40 – 13:00
Isabella Poggi, Francesca D'Errico and Laura Vincze
Ridiculization in Public Debates: Making Fun of the Other as a Discrediting Move

13:00 – 14:00
LUNCH BREAK

ORAL SESSION 3: EMOTION AND AFFECT
(Chair: Laurence Devillers)

14:00 – 14:20
Rene Altrov and Hille Pajupuu
Estonian Emotional Speech Corpus: Theoretical Base and Implementation

14:20 – 14:40
Dipankar Das, Soujanya Poria, Chandra Mohan Dasari and Sivaji Bandyopadhyay
Building Resources for Multilingual Affect Analysis – A Case Study on Hindi, Bengali and Telugu

14:40 – 15:00
Clément Chastagnol and Laurence Devillers
Collecting Spontaneous Emotional Data for a Social Assistive Robot

15:00 – 15:20
Wenjing Han, Haifeng Li, Lin Ma, Xiaopeng Zhang and Björn Schuller
A Ranking-based Emotion Annotation Scheme and Real-life Speech Database

15:20 – 15:40
John Snel, Alexey Tarasov, Charlie Cullen and Sarah Jane Delany
A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora

15:40 – 16:00
Serkan Özkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez and Engin Erzin
Multimodal Analysis of Upper-Body Gestures, Facial Expressions and Speech

16:00 – 16:30
COFFEE BREAK

ORAL SESSION 4: CROSS-DISCIPLINE PERSPECTIVES
(Chair: Haifeng Li)

16:30 – 16:50
Dietmar Rösner, Jörg Frommer, Rico Andrich, Rafael Friesen, Matthias Haase, Manuela Kunze, Julia Lange and Mirko Otto
LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing

16:50 – 17:10
Katia Lida Kermanidis
Mining Authors' Personality Traits from Modern Greek Spontaneous Text

17:10 – 17:30
Antonio Reyes and Paolo Rosso
Building Corpora for Figurative Language Processing: The Case of Irony Detection

17:30 – 17:50
Marcela Charfuelan and Marc Schröder
Correlation Analysis of Sentiment Analysis Scores and Acoustic Features in Audiobook Narratives

17:50 – 18:10
Effie Mouka, Voula Giouli, Aggeliki Fotopoulou and Ioannis E. Saridakis
Opinion and Emotion in Movies: a Modular Perspective to Annotation

18:10 – 18:30
Closing Discussion
Paolo Rosso

19:00 – 21:00
Optional Common Dinner
Covered by participants

Editors

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Organizers/Organizing Committee

Laurence Devillers	Université Paris-Sorbonne 4, France
Björn Schuller	Technische Universität München, Germany
Anton Batliner	Friedrich-Alexander-University, Germany
Paolo Rosso	Universitat Politècnica de Valencia, Spain
Ellen Douglas-Cowie	Queen's University Belfast, UK
Roddy Cowie	Queen's University Belfast, UK
Catherine Pelachaud	CNRS - LTCI, France

Workshop Programme Committee

Vered Aharonson	AFEKA, Israel
Alexandra Balahur	EC’s Joint Research Centre, Italy
Felix Burkhardt	Deutsche Telekom, Germany
Carlos Busso	University of Texas at Dallas, USA
Rafael Calvo	University of Sydney, Australia
Erik Cambria	National University Singapore, Singapore
Antonio Camurri	University of Genova, Italy
Mohamed Chetouani	Université Paris 6, France
Thierry Dutoit	University of Mons, Belgium
Julien Epps	University of New South Wales, Australia
Anna Esposito	IIASS, Italy
Hatice Gunes	Queen Mary University, UK
Catherine Havasi	MIT Media Lab, USA
Bing Liu	University of Illinois at Chicago, USA
Florian Metze	Carnegie Mellon University, USA
Shrikanth Narayanan	University of Southern California, USA
Maja Pantic	Imperial College London, UK
Antonio Reyes	Universidad Politècnica de Valencia, Spain
Fabien Ringeval	Université de Fribourg, Switzerland
Peter Robinson	University of Cambridge, UK
Florian Schiel	LMU, Germany
Jianhua Tao	Chinese Academy of Sciences, China
José A. Troyano	Universidad de Sevilla, Spain
Tony Veale	University College Dublin, Ireland
Alessandro Vinciarelli	University of Glasgow, UK
Haixun Wang	Microsoft Research Asia, China

Table of contents

CROSS-DISCIPLINE PERSPECTIVES

Author Index

Preface

The fourth instalment of the workshop series on Corpora for Research on Emotion held at LREC aims at further cross-fertilisation between the highly related communities of emotion and affect processing based on acoustics of the speech signal, and linguistic analysis of spoken and written text, i.e., the field of sentiment analysis including figurative languages such as irony, sarcasm, satire, metaphor, parody, etc. At the same time, the workshop opens up for the emerging field of behavioural and social signal processing including signals such as laughs, smiles, sighs, hesitations, consents, etc. Besides data from human-system interaction, dyadic and human-to-human data, its labelling and suited models as well as benchmark analysis and evaluation results on suited and relevant corpora were invited. By this, we aim at bridging between these larger and highly connected fields: Emotion and sentiment are part of social communication, and social signals are highly relevant in helping to better understand affective behaviour and its context. For example, understanding of a subject's personality is needed to make better sense of observed emotional patterns. At the same time, non-linguistic behaviour such as laughter and linguistic analysis can give further insight into the state or personality trait of the subject.

All these fields further share a unique trait: Genuine emotion, sentiment and social signals are hard to collect, ambiguous to annotate, and tricky to distribute due to privacy reasons. In addition, the few available corpora suffer from a number of issues owing to the peculiarity of these young and emerging fields: As in no related task, different forms of modelling exist, and ground truth is never solid due to the often highly different perception of the mostly very few annotators. Due to data sparseness, cross-validation without strict partitioning including development sets and without strict separation of speakers and subjects throughout partitioning are frequently seen.

Laurence Devillers, Björn Schuller, Anton Batliner, Paolo Rosso,
Ellen Douglas-Cowie, Roddy Cowie, Catherine Pelachaud

Annotating Irony in a Novel Italian Corpus for Sentiment Analysis

Andrea Gianti*, Cristina Bosco*, Viviana Patti*, Andrea Bolioli◇, Luigi Di Caro*

*Dipartimento di Informatica, Università di Torino
Corso Svizzera, 195, 10149, Torino
andreagianti@gmail.com, {bosco,dicaro,patti}@di.unito.it

◇CELI srl Torino
Via San Quintino 31, 10121, Torino
abolioli@celi.it

Abstract

In this paper we describe our current work on Senti–TUT, a novel Italian corpus for sentiment analysis. This resource includes annotations concerning both sentiment and morpho-syntax, in order to make available several possibilities of further exploitation related to sentiment analysis. For what concerns the annotation at sentiment level, we focus on irony and we selected therefore texts on politics from a social media, namely Twitter, where irony is usually applied by humans. Our aim is to add a new sentiment dimension, which explicitly accounts for irony, to a sentiment analysis classification framework based on polarity annotation.
The paper describes the data set, the features of the annotation both at sentiment and morpho-syntactic level, the procedures and tools applied in the annotation process. Finally, it shows the preliminary experiments we are carrying on in order to validate the annotation work.

Keywords: Irony, Sentiment analysis, Corpus annotation, Social media, Italian

1. Introduction and Motivation

In this paper we describe an ongoing project for the development of an Italian corpus annotated for sentiment analysis. We concentrate our attention on irony, a hard nut that is still to be cracked in the sentiment analysis context, and on a specific topic for texts where irony is usually applied by humans: politics.
Irony is recognized in literature as a specific phenomenon which can harm sentiment analysis and opinion mining systems (Pang and Lee, 2008; Davidov et al., 2011; Tsur et al., 2010). The rhetorical tradition treated irony as the figure of speech in which the meaning is the opposite of the literal meaning, so that an ironists primary intention is to communicate the opposite of what he/she says. Modern Gricean pragmatic theory has not departed radically from this view (Grice, 1975). Another interesting account of irony, the one proposed within relevance theory (Sperber and Wilson, 1986), suggests that irony is a variety of echoic use of language. This approach accounts for cases of "echoic irony", where ironical utterances can be viewed as echoic mentions, in which usually the communicator dissociates herself from the opinion echoed.
The literature on irony and its interpretation is very extensive, however most of the proposals aim at explaining the fact that in an ironic sentence the explicit meaning is different or opposite from the real intended meaning. Therefore, in a sentiment analysis setting the presence of ironic devices in a text can work as an unexpected "polarity reverser", by undermining the accuracy of the systems, especially in application contexts focussing on monitoring political sentiment, where blogs or social media provide the data sources. Recently, such application contexts gained popularity, since message content from social media (microblogging like Twitter[1] especially) turned out to be a powerful real-time indicator of political sentiment. Microblogging messages, like "tweets" or Facebook messages, emerged as a very valuable information data not only in politics, but in a variety of NLP application domains, ranging from the extraction of critical information during times of mass emergency (Verma et al., 2011) to the sentiment analysis for the stock market prediction (Bollen et al., 2010).
However, Twitter communications includes a high percentage of ironic and sarcastic messages (Davidov et al., 2011; Tumasjan et al., 2011), and platforms monitoring the sentiment in Twitter messages experimented the problem to classify as positive many posts which instead express ironic non-positive judgments or opinions. As an example, let us consider the following tweet [2]:

TWSPINO-1160
'Alemanno: "Questa mattina sembra tutto funzionante". Gli hanno spiegato come funziona la pala'
(Alemanno: "This morning everything seems to be working properly." They've showed him how the shovel works)

In absence of irony recognition, such tweet it is classified as positive, while it clearly expresses a criticism w.r.t. the Rome's mayor ability to deal with the snow emergency in Winter 2011-2012[3].
In our tweets, we observed the presence of the well-known lexical devices and features that characterize humorous

[1] http://twitter.com

[2] In february 2012, Rome's mayor, Gianni Alemanno, was widely criticised in Italy for failing to activate an emergency plan after an exceptionally heavy snowfall.

[3] English translations of the Italian examples are mainly literal and so may sometimes appear awkward in English.

texts, like linguistic ambiguity, the use of affective terms, and so on, i.e. the tweet TWSPINO-32: *'Marchionne presenta la nuova Panda. Il timore è che si diffonda tra la popolazione'* (Marchionne has presented the new Panda. It is feared that it may spread throughout the population).[4]
Moreover, we observed many cases of "echoic mentions" (Sperber and Wilson, 1986) among our ironic tweets. For instance in tweet TWNEWS-570 *'Governo Monti: la rassicurante conferma che in Italia non esistono Tecnocrati, che non siano Gerontocrati. Non è un Paese per giovani'* (Monti's government: the reassuring confirmation that in Italy do not exist Technocrats which are not Gerontocrats. No country for young men.) the sentence *'non è un paese per giovani*' (no country for young men) is a case of echoic mention, with a clear reference to the title of the movie *'Non è un Paese per Vecchi'* (No Country for Old Men[5]).
The main aim of this project is to add a new sentiment dimension, which explicitly accounts for irony, to a sentiment analysis classification framework based on polarity annotation. To the best of our knowledge, existing sentiment analysis frameworks consider the following dimensions: subjectivity and objectivity; (positive or negative) polarity; emotional categories; opinions about entities. Accordingly, corpora that are manually annotated for subjectivity, polarity, or emotion, are available in many languages. Nowadays, with few exceptions (Esuli et al., 2008), Italian is among the less-resourced languages with respect to sentiment analysis. For what concerns English, let us mention the MPQA Opinion Corpus[6], which contains news articles from a wide variety of news sources manually annotated for opinions and other private states (like emotions, sentiments, etc.). A multilingual dataset[7], automatically annotated for subjectivity, in English, Arabic, French, German, Romanian, and Spanish, is the result of the work described in (Banea et al., 2010), while the multilingual corpus (Spanish, Italian and English) of blog posts in (Boldrini et al., 2010) is annotated according to the *EmotiBlog* annotation schema.
In the last years the authors gained experience both in sentiment analysis applied to social media (CELI and Me-Source, 2009), and in ontology-driven sentiment analysis applied to socially tagged resources (Baldoni et al., 2012), with a focus on the Italian language. Moreover, some among them are actively involved from more than ten years in both the development of linguistic resources morphosyntactically annotated, namely the treebank TUT (Bosco et al., 2000) (see Section 2.2.), and the exploitation of annotated data in several contexts for training and evaluation of NLP tools, see e.g. (Bosco and Mazzei, 2012b) and (Bosco and Mazzei, 2012a). On this line, we are now working to make available a novel Italian corpus for sentiment analysis, that we call Senti–TUT, which includes sentiment annotations concerning irony and consists in a collection of texts from social media. Such kind of resource is currently missing in particular for Italian. Moreover, we are carrying on some preliminary experiments in classification of our data in order to validate the annotation work.
The paper is organized as follows. In the next section we describe the corpus and the annotation we applied on it. Then, we discuss the preliminary experiments performed for the validation of data. The last section outlines some directions for future work.

2. Data

In this section we describe the data collected for the Senti–TUT project and the annotation we are applying on them. All the data related to the project and the information about download can be found in the Senti–TUT web site: http://www.di.unito.it/˜tutreeb/sentitut.html.

2.1. The corpus

As confirmed by various references (Davidov et al., 2011) and (Tumasjan et al., 2011) social media, such as Facebook or Twitter, includes a high percentage of ironic and sarcastic messages and can mirror offline political sentiment, as they did for instance in the recent USA and German elections. Our linguistic data are therefore mainly collected by Twitter.
As far as the text style is concerned, in general, Twitter communications are composed by messages called "tweets", each of which is shorter than 140 characters and can be composed by one or more sentences. In our Italian corpus of messages most of tweets are composed by two short sentences or simple noun phrases, and very rarely by wh-sentences. The typical structure of a tweet is shown in the following post[8]:

TWSPINO-107
'Napolitano: "Attenti a toccare la Costituzione". Bisogna aspettare il medico legale.'
(Napolitano: "Be careful you don't touch the Constitution". We have to wait for the forensic surgeon to arrive first.)

With respect to the composition and size of the data set, it is organized in two subcorpora, namely TWNEWS and TWSPINO. The former is currently composed of around three thousands of tweets, published in the weeks after the new Italian prime minister Mario Monti announced his Cabinet (from October 2011 the 16th to February 2012 the third). The latter is instead composed of more than one thousand tweets extracted from the Twitter section of Spinoza, published from July 2009 to February 2012. Spinoza[9], is a very popular collective Italian blog which includes a high percentage of posts with sharp satire on politics, which is published on Twitter since 2009. This subcorpus has been therefore added in order to enlarge our data set with texts where various forms of irony are involved. The collection of all the data has been done by exploiting a collaborative annotation tool, which is part of the Blogmeter social media monitoring platform (CELI and Me-Source,

[4] Marchionne is CEO of the Italian automotive group Fiat. Panda is the name of a Fiat city car.
[5] For details, see the Wikipedia page: http://en.wikipedia.org/wiki/No_Country_for_Old_Men_(film).
[6] http://www.cs.pitt.edu/mpqa/
[7] http://www.cse.unt.edu/˜rada/downloads.html#msa
[8] Giorgio Napolitano is the current President of the Italian Republic.
[9] http://www.spinoza.it/

```
1 La (IL ART DEF F SING) [7;VERB-SUBJ]
2 spazzatura (SPAZZATURA NOUN COMMON F SING) [1;DET+DEF-ARG]
3 di (DI PREP MONO) [2;PREP-RMOD]
4 Napoli (NAPOLI NOUN PROPER F SING ££CITY) [3;PREP-ARG]
5 si (SI PRON REFL-IMPERS ALLVAL ALLVAL 3 LSUBJ+LOBJ+LIOBJ CLITIC) [7;VERB-OBJ]
6 sta (STARE VERB AUX IND PRES INTRANS 3 SING) [7;AUX]
7 decomponendo (DECOMPORRE VERB MAIN GERUND PRES TRANS) [0;TOP-VERB]
8 . (#\. PUNCT) [7;END]

1 Concorrerà (CONCORRERE VERB MAIN IND FUT INTRANS 3 SING) [0;TOP-VERB]
1.10 t [] (GENERIC-T PRON PERS ALLVAL ALLVAL ALLVAL) [1;VERB-SUBJ]
2 al (A PREP MONO) [1;VERB-INDCOMPL]
2.1 al (IL ART DEF M SING) [2;PREP-ARG]
3 Nobel (NOBEL NOUN PROPER) [2.1;DET+DEF-ARG]
4 per (PER PREP MONO) [3;PREP-RMOD]
5 la (IL ART DEF F SING) [4;PREP-ARG]
6 chimica (CHIMICA NOUN COMMON F SING) [5;DET+DEF-ARG]
7 . (#\. PUNCT) [1;END]
```

Figure 1: The tweet 216 from the Spinoza corpus (TWSPINO-216) as annotated in TUT format.

2009). These data are only a portion of the whole material collected by this tool for the above mentioned periods (which are about 11,000 tweets).

2.2. The annotation

The project for the development of the Senti–TUT involves the annotation of the linguistic data with respect to two distinguished levels. While the first one includes morphological and syntactic tags as usual e.g. in treebanks, the second refers instead to concepts typical of sentiment analysis.

2.2.1. Morphological and syntactic annotation

For what concerns the morphological and syntactic annotation, this is done according to the format developed and applied in the Turin University Treebank (henceforth TUT) project (Bosco et al., 2000). This treebank is a freely available resource developed by the Natural Language Processing group of the University of Turin (for more details and examples see http://www.di.unito.it/˜tutreeb) including 102,150 annotated tokens (around 3,500 sentences), which has been successfully exploited as testbed in various evaluation campaigns for Italian parsing (http://www.evalita.it/, (Bosco and Mazzei, 2012b) and (Bosco and Mazzei, 2012a)). We selected this format for two main reasons: the reliability of TUT format for the involved language and the availability of a variety of tools implemented within TUT project, first of all the Turin University Linguistic Environment (TULE, http://www.tule.di.unito.it/, (Lesmo, 2007) and (Lesmo, 2009)), whose pipeline includes tokenization, morphological and syntactic analysis.

In figure 1 and 2, a post extracted from our tweet corpus is represented according to TUT format: TWSPINO-216 *'La spazzatura di Napoli si sta decomponendo. Concorrerà al Nobel per la chimica.'* (The garbage of Naples is becoming rotten. It will apply for the chemistry Nobel prize.). In particular, we can observe that TUT format is featured by a very detailed morphological tag set, which is useful for the description of a language with a rich inflection, and by a large inventory of grammatical relations exploited in the labeling of the edges of the dependency trees. For each word, the lemma, the morphological category and related features are annotated together with the index of the father in the dependency tree and the relation linking the word with the father itself. Moreover, in order to offer an explicit representation of all the elements involved in the predicate argument structure, e.g. the subject which is often dropped in Italian, TUT format includes also null elements, see e.g. the annotation of the node 1.10 (t) which is the subject of the second sentence of the tweet represented in the figures. The morpho-syntactic annotation of the Senti–TUT corpus is automatically performed by TULE and then semi-automatically corrected by exploiting the tools developed within the TUT project. Nevertheless, the application of these tools, TULE especially, to the Senti–TUT corpus shows that, in order to achieve reliable annotations, the integration in the parsing process of various patterns typical of the social media language is needed. These patterns vary from the use of several citations from the Web to the words and phrases not formal or literary. Twitter, and social media in general, represent in fact a text genre different from those previously analyzed by exploiting TULE, e.g. newspaper or legal, which has never been analyzed in our knowledge for Italian. It is known in literature that in order to obtain a reliable morphological and syntactic analysis of a specific text genre, the parsing systems should be carefully tuned on the basis of it (Gildea, 2001). This is clearly showed by the current performance scores of TULE parser, which are far from those obtained on the text genres included in TUT, in particular with respect to the syntactic analysis. Nevertheless, the Evalita experiences showed evidences that TULE and other parsing systems for Italian can achieve, if trained and tuned, performances close to the state of the art for English for various text genres.

2.2.2. Annotation for sentiment analysis

As far as the annotation at the level useful for sentiment analysis is concerned, the data are currently annotated at tweet level, since one sentiment tag is applied to each tweet (considering that a tweet can be composed by more than one sentence). Nevertheless, even if, for the present time,

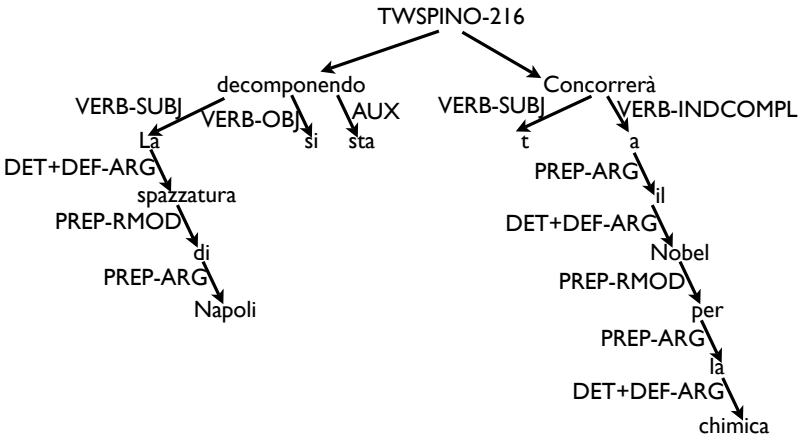

Figure 2: The TUT dependency tree for the tweet 216 from the Spinoza corpus (TWSPINO-216).

the focus of the Senti–TUT is mainly the annotation at tweet level, the resource we are currently developing has to be seen in the wider framework of a project for sentiment analysis and opinion mining. And within this context it should be considered also the availability of the morpho-syntactic annotation on the same data, which allows in the future for the application of other more fine-grained annotations and analysis related to sentiment analysis. For instance, the availability of Part of Speech tags and lemmas for words allows for investigations that relate morphological and sentiment features, e.g. adjective which are carried on sentimental meaning. As in (Tsur et al., 2010) syntactic features can be useful in the identification of irony, e.g. the use of punctuation.
In the table below the sentiment tags used for the annotation of Senti–TUT are described.

Sentiment tag	Meaning
POS	positive
NEG	negative
HUM	ironic
NONE	objective (none of the above)
MIXED	POS and NEG both

Table 1: The sentiment tags applied in Senti–TUT.

The following are examples of the annotation of tweets with the above mentioned sentiment tags.

TWSPINO-30 (tagged as HUM)
'C'e' cosi' tanta crisi che Babbo Natale invece delle letterine riceve curriculum.'
(The economic crisis is so hard that Santa Claus receives curricula vitae instead of letters.)

TWNEWS-123 (tagged as NONE)
'Mario Monti premier? Tutte le indiscrezioni.'
(Mario Monti premier? All the gossips.)

TWNEWS-24 (tagged as POS)
'Marc Lazar: "Napolitano? L'Europa lo ammira. Mario Monti? Puo' salvare l'Italia"'
(Marc Lazar: "Napolitano? Europe admires him. Mario Monti? He can save Italy")

TWNEWS-124 (tagged as NEG)
'Monti e' un uomo dei poteri che stanno affondando il nostro paese.'
(Monti is a man of the powers that are sinking our country.)

TWNEWS (tagged as MIXED)
'Brindo alle dimissioni di Berlusconi ma sul governo Monti non mi faccio illusioni'
(I drink a toast to the Berlusconi's resignation, but
I have no illusion about the Monti's government)

We also used the tag UN in order to mark tweets which are not classifiable, e.g. tweets containing incomplete or meaningless sentences, which are therefore discarded. The distribution of the tags can be seen by observing a preliminary data set composed by around 1,500 tweets: around a third is classified as NONE, 400 as NEG, 300 as HUM, 250 as POS, and the remaining as MIXED or UN.
While the morpho-syntactic annotation is automatically performed by TULE, the annotation of the sentiment tags at the tweet level is currently manually performed by exploiting a collaborative annotation tool, which is part of the Blogmeter social media monitoring platform (CELI and Me-Source, 2009). Among the utilities made available by Blogmeter we applied, in particular, those related to filtering out the non relevant data, e.g. the *re-tweets* (i.e. the forwarded tweets).
Five human skilled annotators have been involved until now in this annotation task producing for each tweet not less than two independent annotations. This manual annotation helped by Blogmeter has been followed by an inter-annotator agreement check, as usual in the development of linguistic resources. In order to solve the disagreement, which can be referred to about 25% of the data, the independent annotation of a third human has been applied to the ambiguous tweets (i.e. those where each of the two annotators selected a tag different from the other annotator). The cases where the disagreement persists (i.e. tweets where

each of the three annotators selected a tag different from the others), which are around 3%, have been then discarded since considered as too ambiguous to be classified.

3. Preliminary experiments

We are carrying on some preliminary experiments in classification of our data in order to validate the annotation work. These experiments are based on a portion of the Senti–TUT corpus and more precisely on about 1,550 annotated tweets from TWNEWS with a balanced tagging of the four above indicated sentiment labels.

Starting from the promising results for other languages (Strapparava et al., 2011; Davidov et al., 2011), we are setting up a framework where irony recognition in our tweets can be formulated as a classification task and machine learning algorithms can be applied.

Making use of a simple evaluation scheme for classification-based tasks called Confusion Matrix (Stehman, 1997), it is possible to look at the existing overlapping among the classes, i.e., how much one class is misclassified as another one. This mechanism usually gives some hint on the lexical overlapping between the texts of two different classes. In our case, we noticed a significant lap between humorous texts and negative ones, while the same does not happen when comparing humorous with positive texts. This somehow confirms what already discovered by (Mihalcea and Pulman, 2007).

Another interesting point of analysis concerns the discriminatory power of the words within the classification procedure. This can be easily done by calculating the Information Gain (or Kullback-Leibler divergence (Kullback and Leibler, 1951)) of the terms with respect to the class labels. In case of comparisons between texts sharing both temporal and domain characteristics, it helps to discover current targets of humor. For instance, using our recent tweets talking about Italian politics, terms like 'Monti' and 'Passera' resulted to be highly relevant during classification (the first one refers to the current italian prime minister Mario Monti, whereas the second is the italian minister of economy and development Corrado Passera). Notice that both 'monti' and 'passera' are words of the Italian vocabulary (e.g. the word "monti' means 'mountains', while 'passera' means 'hen sparrow' but it is also used in adult slang as masculinist metaphor), and many jokes in our tweets exploit such forms of ambiguity.

As a second result, this tool allows to individuate those recurrent patterns that are strictly related to the information sources. In our scenario, the token "http" usually indicates the presence of news instead of humorous texts. This is due to the shortness nature of Twitter that obliges the users to be concise. Indeed, most of non-humorous and informative tweets contain few words followed by one hyperlink (e.g TWNEWS-186: *'Chi è Mario Monti? http://t.co/BZewchzZ'* (Who is Mario Monti? http://t.co/BZewchzZ).

Still, Information Gain can be used to mine those linguistic expressions, rather than single words, that can be useful to identify the humorous nature of the text. For example, meaningful terms that turn out to be important in this sense are "speriamo" (i.e., "we wish") and "bene"("good"), which refer to the italian expression "speriamo bene" ("fingers crossed"). Other highly-scored terms include "fiducia" ("trust"), "finalmente" (finally), and so forth. One next step in this direction would be to evaluate such discriminatory power with respect to each one of the classes.

In future works, we aim at using linguistic resources to preprocess the input texts in order to remove noise and uninformative terms. Then, the use of data morpho-syntactically annotated could be crucial in the identification of whole syntactic structures (e.g., "bank director") as well as linguistic expressions. Finally, the time and the mood of verbs can be another way of studying linguistic differences between humorous and objective texts.

All the above points only represent some issues that came out from our first experiments, thus they are to be considered as preliminary results.

4. Conclusion and future work

In this paper we described our current work on Senti–TUT, a novel Italian corpus for sentiment analysis which includes sentiment annotations concerning irony and consists in a collection of texts from Twitter.

For what concerns issues arising in the manual annotation of the sentiment of our tweets, useful guidelines were found in (Wiebe et al., 2005), where a general annotation scheme to distinguish subjective information from material presented as fact is defined. Tweets in our corpus often express opinion about news entities while reporting on recent events (Godbole et al., 2007), or report opinions of news entities (e.g. politicians) about the breaking news. Following (Wiebe et al., 2005) in both cases we considered the tweets as subjective (with a positive or negative polarity).

Concerning the specific issue of determining if a tweet is ironic, this is not an easy task, mainly due to the fact that irony is very subjective and personal appreciation can lead to different perceptions. We mainly recognized the following features in our tweets: frequent use of adult slang and dirty words, use of echoic irony, language jokes, which often exploit ambiguities involving the politicians' proper nouns, as confirmed by first experiments. Moreover, we observed many cases of quotation or explicit reference to popular, Italian or international, television series, see e.g. the following tweet referring to the American reality television series Jersey Shore: TWNEWS-844 '*@mtvitaly ma è vero che Mario Monti parteciperà a Jersey Shore? http://t.co/d0H1Kmp6'* (@mtvitaly Is it true that Mario Monti will be a cast member of Jersey Shore? http://t.co/d0H1Kmp6). Therefore, a problem that needs to be taken into account is that sometimes in our context the recognition of irony can be hard, because strongly depends not only to the annotator knowledge about the Italian political situation but also to his/her degree of "addiction" to tv shows.

Since the perception of irony can vary from a subject to another, different annotators could consider a given post ironic or sarcastic "to some degree". In order to face this issue, it would be useful to assign scores to ironic annotations, as suggested in (Davidov et al., 2011). Moreover, we are also considering to extend the annotation framework by adding a more fine-grained annotation where the entire

text is divided in pieces (or fragments) representing both the facts under discussion and the expressions about the judgement. In such richer setting, it will be possible to evaluate the system at different levels of granularity and to use the information to measure different degree of irony. Moreover, during the annotation work, we have observed many different typologies of ironic statements, as for instance *sarcastic tweets*, conveying bitter or cutting expressions or remarks, *hilarious or facetious tweets*, aimed at producing a comic effect, *language jokes*, and so on. In order to tackle this issue, as future work we aim at studying a more sophisticated classification of ironic tweets, where different ways of expressing irony can be distinguished (and possibly organized in a taxonomy) and tweets can be annotated accordingly. In this framework it will be also interesting to test the results of enabling multi-value-annotations.

5. Acknowledgements

The work reported in this paper was partly carried out in the framework of the project PARLI ("Portal for the Access to the Linguistic Resources for Italian") benefitting from a research funding PRIN-2008 from the Ministry of Public Instruction and University.
The authors thank all the persons who supported the work. We are grateful to our annotators, in particular to Gianna Cernuschi and Andrea Marchetti and to CELI Torino for providing the facilities offered by the Blogmeter social media monitoring platform.

6. References

Matteo Baldoni, Cristina Baroglio, Viviana Patti, and Paolo Rena. 2012. From Tags to Emotions: Ontology-driven Sentiment Analysis in the Social Semantic Web. *Intelligenza Artificiale: the International Journal of the AI*IA*. In press.

Carmen Banea, Rada Mihalcea, and Janyce Wiebe. 2010. Multilingual subjectivity: are more languages better? In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 28–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ester Boldrini, Alexandra Balahur, Patricio Martínez-Barco, and Andrés Montoyo. 2010. Emotiblog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Cristina Bosco and Alessandro Mazzei. 2012a. The evalita 2011 parsing task: the constituency track. In *Evalita'11 Working Notes*, Roma.

Cristina Bosco and Alessandro Mazzei. 2012b. The evalita 2011 parsing task: the dependency track. In *Evalita'11 Working Notes*, Roma.

Cristina Bosco, Vincenzo Lombardo, Leonardo Lesmo, and Daniela Vassallo. 2000. Building a treebank for Italian: a data-driven annotation schema. In *Proceedings of LREC'00*.

CELI and Me-Source. 2009. Blogmeter. http://www.blogmeter.eu/.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2011. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.

Andrea Esuli, Fabrizio Sebastiani, and Ilaria Urciuoli. 2008. Annotating expressions of opinion and emotion in the italian content annotation bank. In *Proceedings of the 6th Conference on Language Resources and Evaluation (LREC 2008)*.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of the Empirical Methods in NLP Conference*, Pittsburg.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 64–75. Academic Press, New York.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

Leonardo Lesmo. 2007. The rule-based parser of the NLP group of the University of Torino. *Intelligenza artificiale*, 2:46–47.

Leonardo Lesmo. 2009. The Turin University Parser at Evalita 2009. In *Proceedings of Evalita'09*.

Rada Mihalcea and Stephen Pulman. 2007. Characterizing humour: An exploration of features in humorous texts. In *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing '07, pages 337–347, Berlin, Heidelberg. Springer-Verlag.

Bo Pang and Lillian Lee. 2008. *Opinion Mining and Sentiment Analysis (Foundations and Trends(R) in Information Retrieval)*. Now Publishers Inc.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: communication and cognition*. Harvard University Press, Cambridge, MA, USA.

Stephen V. Stehman. 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77 – 89.

Carlo Strapparava, Oliviero Stock, and Rada Mihalcea. 2011. Computational humour. In Roddy Cowie, Catherine Pelachaud, and Paolo Petta, editors, *Emotion-Oriented Systems*, Cognitive Technologies, pages 609–634. Springer Berlin Heidelberg.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm - a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In William W. Cohen and Samuel Gosling, editors, *In Proceedings of the Fourth International Conference on Weblogs and Social Media (ICWSM-2010)*. The AAAI Press.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2011. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 178–185. AAAI.

Sudha Verma, Sarah Vieweg, William J. Corvey, Leysia Palen, James H. Martin, Martha Palmer, Aaron Schram, and Kenneth M. Anderson. 2011. Natural language processing to the rescue? extracting situational awareness tweets during mass emergency. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 385–392. AAAI.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Sentiment analysis resources for Slovene language

Sandi Pohorec, Ines Ceh, Peter Kokol, Milan Zorman
University of Maribor, Faculty of electrical engineering and computer science, Smetanova ulica 17, SI-2000 Maribor, Slovenia
{sandi.pohorec, ines.ceh, kokol, milan.zorman}@uni-mb.si

Abstract

Slovene language lacks resources for sentiment analysis of natural language. Several large lexical resources are available, but they only provide information on word lemmas and part-of-speech tags. This paper presents an experiment in which the well-known General Inquirer (GI) dictionary has been automatically translated into Slovene with the use of several multilingual resources. We have implemented an automated system for the translation of the General Inquirer dictionary as well as processed large amounts of Slovene text in order to provide the basic statistical data, used for language recognition, in the form of n-gram distributions. Each word entry in the translated dictionary has been lemmatized and each entry provides all known word forms. The resources presented here offer the capability to automatically detect if the text is in Slovene language and analyze the content with GI regardless of the word form.

1. Introduction

Slovene is a highly inflectional language. Along with Croatian, Serb, Macedonian and Bulgarian it forms the south Slavic language group. Slovenia is the area where the Slavic languages meet with Romanian, Germanic and Finno-Ugric languages; consequently Slovene is a language with many specific characteristics in phonology, lexicology and morphology. Some very important lexical resources for Slovene have been developed in the past, focusing mostly on the part-of-speech tagging and lemmatization. Jos (Erjavec et al., 2010; Erjavec & Krek, 2008) is a validated linguistically and morphosyntactically tagged corpora. The 100.000 word version of the corpora has been annotated manually. A major resource is also available in the form of MULTEXT-East (Erjavec, 2010). Multext-East is a multilingual, (Bulgarian, Croatian, Czech, Estonian, English, Hungarian, Romanian, Serbian, Slovene, Resian dialect of Slovene, Macedonian, Persian, Polish, Russian, Slovak, and Ukrainian) standardized and linked set of morphosyntactic specifications; morphosyntactic lexica; and annotated parallel, comparable, and speech corpora. Slovene WordNet (Erjavec & Fišer, 2006) offers a lexical database (approximately 5000 top-level concepts) which organizes nouns, verbs, adjectives and adverbs in conceptual hierarchies, thereby linking semantically and lexically related concepts.

The available resources however offer no significant value for sentimental analysis of Slovene natural language. This paper focuses on an experiment to automatically translate the GI dictionary to Slovene. To accompany the translated dictionary a large amount of Slovene corpora have been processed to provide n-gram occurrence frequencies for the language. These can be used to automatically verify if the language of the processed content is in fact in Slovene. The paper is organized as follows: Section 2 briefly introduces related work in sentiment analysis; Section 3 introduces the GI dictionary, the size of the dictionary and the categories it contains. Section 4 discusses the process of the translation. In Section 5 we briefly introduce the process of acquiring the n-gram occurrence frequencies which can be utilized for automated language recognition. Automated detection of language of analyzed documents can accommodate the use of sentiment resources in multiple languages because it can render the automated resource selection according to the target language. Section 6 is dedicated to the evaluation of the accuracy of the translation and the applicability of the dictionary on unknown content. The paper ends with concluding remarks in Section 7.

2. Sentiment analysis

Sentiment analysis (SA) is aimed at the identification of opinions, emotions and evaluations expressed in natural language (Wiebe, 1994; Thet et al., 2010). Sentiment is the deviation from neutral orientation of subject discourse. Sentiment is classified as positive or negative. The target of the sentiment is the object/subject that the sentiment in the text is aimed at. Major goal of research is the automated determination of sentiment orientation or polarity (negative, neutral, positive) of analyzed text. The analysis is done on individual words, phrases, sentences or paragraphs of analyzed text. SA depends on lexical resources to identify sentiment bearing words and determine their polarity. General Inquirer dictionary, created at Harvard (Stone and Hunt, 1963) is a manually created resource, often used in SA research.

(Hatzivassiloglou and McKeown, 1997) used a machine learning approach to construct a lexicon of sentiment terms. Multiple techniques and approaches have been proposed for the identification of word polarity (Thet et al., 2010): extraction of adjectives (Turney, 2002; Wiebe 2000), nouns (Riloff et al., 2003), and linguistics patterns from subjective expressions (Riloff and Wiebe, 2003). A propagation approach for extracting large number of sentiment words with assigned polarity was proposed (Qiu et al., 2009). Support vector machines were proven to perform better than naïve Bayes and maximum entropy classification (Pang et al., 2002) when assigning document polarity. (Mullen and Collier 2004) introduced a hybrid of SVM approach combined with favorability

measures of terms and topics. Measures of favorability of terms and topic polarity inherently rely on resources. Several resources are currently available, among them: *Dictionary of Affect of Language* (DAL; Whissell, 1984), *WordNet* (Miller, 1990) and a more recent SentiWordNet (SWN) 3.0 (Baccianella, 2010). DAL is a dictionary of 8742 manually rated words with respect to their activation evaluation and imagery. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into cognitive synonyms, each group expressing a distinct concept. SWN is a lexical resource for opinion mining, in SWN each cognitive synonym from WordNet is assigned with three sentiment scores: positivity, negativity and objectivity.

3. General Inquirer dictionary

General Inquirer (GI) is a dictionary of tag categories: (1) Harvard IV-4 dictionary, (2) Lasswell value dictionary, (3) several recently constructed categories and (4) marker categories. Marker categories are primarily used for disambiguation. The dictionary is a set of 11788 words annotated according to a set of 182 tag categories. Each category is a list of words and word senses. GI categories have been developed (manually) for social science content-analysis. Two large valence categories tag 1915 words as being positive and 2291 as being negative (negative is the largest category in the dictionary). Other 180 categories annotate words by pleasure, pain, virtue, vice, words of overstatement or understatement, usage by institution, words for roles, rituals, relations etc. The main processing involved with the usage of the dictionary is the disambiguation process. In order to achieve reasonable accuracy the correct sense has to be assigned to the word being processed (on the condition that the word is a homonym). GI removes common regular suffixes so that one entry can match multiple inflected word forms. Each entry is either a root of a word, inflected form of a word or a (multiple entries) word sense. The routines that stem words, along with dictionaries and disambiguation routines limit the general applicability of the dictionary only to the English language.

4. Slovene GI

The translation of the dictionary was set to be as automated as possible. Therefore we have assembled a large amount of Slovene corpora (Erjavec and Fišer 2006; Erjavec and Krek 2008) as well as bilingual dictionaries (Erjavec 2010). The process of translation started with creating a list of all the dictionary entries. Each entry was additionally marked if it has multiple senses (homonym words). Homonyms are marked in the original GI with a number sign (#) and a consequent number. For instance the word "thank" has four possible senses:

- expression of gratitude (verb),
- acknowledgement of gratitude (noun),
- idiom-interjection ("thanks") and
- idiom-noun ("thanks to").

Analysis of homonym words shoved that there were 1603 homonyms, with additional 3147 senses therefore 4750 words out of 11788 (40.3%) were words that required more than just reliable translation. Other 7038 words have been annotated with a single sense of the word therefore no mapping by sense was required. The process of the translation was done in two phases: a) the translation of single sense words and b) the translation of multi sense words.

4.1 Translation of single-sense words

The translation of single sense words was done by querying all the resources available for the translation of each individual word. Then a decision on the choice for the translation was done by voting. The translation resources in our approach are grouped into an ensemble. Each member is weighted by a confidence factor α_k. The translation is chosen with the use by selecting the translation with the highest score. The score is calculated as follows: $s(t_n) = \sum_{k=1}^{K} \alpha_k * r(t_n)$, where t_n is the n-th translation, K is the number of translation resources, $r(t_n)$ is a binary function where the value is 1 if the resource suggests this translation and 0 if the resource suggests another translation.

This is a pretty straightforward process; however the possibility, that a word that was marked as a single sense word in GI would be translated into a word with multiple senses in Slovene has to be considered. This represents a challenge because single sense words do not have a description (in the GI) of the sense of the word. With no description that would provide the exact meaning of the word there is no method available to automatically map the single English entry to one of the possible Slovene translations. Therefore currently only the words, where there is no ambiguity of the Slovene sense, are automatically translated.

4.2 Translation of multi-sense words

Multi-sense words in the GI are additionally annotated by each individual sense. Each annotation gives a percentage value of how common that sense is. For example the word "thank" has 4 possible senses, each as a different part-of-speech (each sense is marked with percentage values of the occurrence of that sense): 49% verb, 6% noun, 37% idiom-interjection, 8% idiom-noun. Following the part-of-speech tag is a short description, "thank" as a noun is described as "to express gratitude, give thanks to". We have used the part-of-speech tag and the short description to match each sense to the Slovene counterpart of that sense. The word "thank" in Slovene is "hvala". However "thank" in the sense of expression of gratitude is "zahvala". The mapping was done with the transformation of the word descriptions in GI and all possible senses of the translated words to a semantic network and making comparisons of the sense of each entry. The process is represented in Fig 1. At the beginning the English word sense with the description and POS tags is read from GI. Then at step one the word is translated to Slovene. In step two the semantic net (Luger,

2005) of the meaning of the English description is generated. In step three the Slovene sense descriptions are translated and in step four they are transformed to semantic net representation. In step five the Slovene and English semantic nets are compared. Best matching individuals are considered an accurate translation.

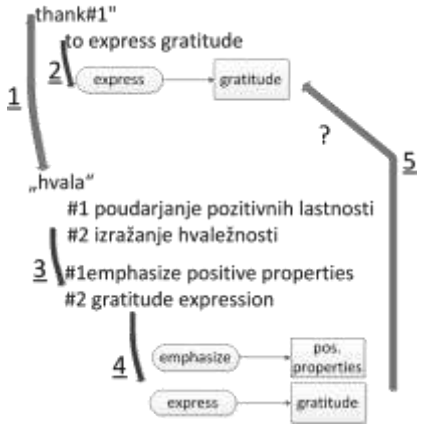

Figure 1: The mapping of a sense of "thank" to Slovene

4.3 Lemmatization and assigning inflected word forms

Major obstacles, that limits the use of GI in other languages, are the associated routines (stemming and/or lemmatization, disambiguation), which have to be duplicated for other languages. We have processed the translated dictionary with the use of freely available Slovene resources and have assigned individual dictionary entries with their basic words and all available inflected forms. Table 1 presents some statistics on the inflected forms of the dictionary entries. The translated dictionary contains 7435 distinct base forms of words (multi sense words count only once). The total of all inflected word forms is 53553, which is an average of 7.2 inflected forms per lemma. The maximum inflected forms per lemma were 24 and there are 66 lemmas for which no inflected forms were added.

5. Character level n-gram language identification

In order to enable sentimental analysis it is essential to be able to automatically detect the source language of processed documents. Language identification was first explored in cryptography, where a *k*-gram, character level language identification algorithm was presented (Konheim, 1981). Several other researchers have been studying language identification and confirmed the n-gram character technique to be successful (Beesley, 1998; Cavnar & Tenkle, 1994).

Although, as we mentioned previously, there are several large scale Slovene language resources available, we have found no statistical data on n-gram distribution. These are trivial to obtain if enough written text is available, but they do require some processing. To obtain valid n-gram occurrence frequencies a large amount of written text is required. We have selected a balanced (by contributing source and topic) set of 3.050.725 distinct Slovene words to calculate n-gram occurrence frequencies. Tables 2 and 3 respectively present the ten most common uni- and bigrams for Slovene. Each table provides a comparison to English unigrams and bigrams although the results for English were acquired on a much smaller data set of 72.540 words, so the values for English should not be fully trusted. The occurrence values have been normalized to show the percentage occurrence for individual n-grams.

Distinct lemmas	7435
Inflected forms	53553
Max inflected words per lemma	24
Lemmas with inflected word forms (min 1)	7369
Lemmas with no inflected word forms	66
Average inflections per lemma	7.2

Table 1. Statistical data on the lemmatization of the Slovene entries of the translated GI

Slovene		English	
a	11.01%	e	10.19%
i	9.63%	i	7.70%
e	8.67%	s	6.97%
o	8.28%	a	6.72%
n	6.69%	r	6.40%
r	6.09%	n	6.37%
t	5.07%	t	6.24%
m	4.77%	o	5.48%
l	4.54%	l	4.90%
s	3.99%	c	3.54%

Table 2. Unigram occurrence statistics (top 10) for Slovene, compared to English

Slovene		English	
ni	1.77%	in	2.30%
ra	1.65%	er	1.87%
st	1.51%	es	1.73%
ne	1.38%	ti	1.54%
ma	1.27%	re	1.38%
em	1.22%	ed	1.38%
ti	1.22%	ng	1.38%
re	1.22%	on	1.38%
im	1.20%	te	1.36%
en	1.19%	at	1.28%

Table 3. Bigram occurrence statistics (top 10) for Slovene, compared to English

6. Evaluation of the applicability of the translated dictionary

In order to verify the newly created Slovene translation of the General Inquirer dictionary we have performed two essential tests. The first test was aimed at validating the correctness of the semi-automatic translation process and was performed manually. The second test was aimed at determining the applicability of the dictionary on a large set of Slovene texts.

Validation of the dictionary (first test; done manually) was performed over a selection of a random sample of words. Approximately one tenth of the dictionary (1000 words, 8,4% of all words) were randomly selected and manually verified if they are the accurate translations of the original entries. The test showed that 7.6% of the words were in fact incorrectly translated. The erroneous words were in two categories: some were not translated at all while others were actually translated into languages other than Slovene (mostly Croatian or Serb). To verify how many words were actually not translated at all we counted all the words where the translation is identical to the source (in the entire dictionary). We have found 588 such words (4.9% of the entire dictionary).

Evaluation of the general applicability of the translated dictionary (second test, performed automatically) was done by processing 140.247 news items. Table 4 shows the size of the data set and the coverage percentage of the dictionary entries. This test was aimed at covering the applicability of the translated dictionary to large scale corpora of Slovene texts. The news items were tokenized (they contained 4.9473.505 words) and each word was checked if it is an entry in the translated dictionary (we have used all forms of the entries). The test shoved that almost 33% (1.628.170 words) in the news items were entries in the translated GI.

Number of news items	140.247
Number of areas the news covered	14 (politics, economy, sport, health, tech….)
Number of words in news items	4.947.505
Number of words covered by the GI	1.628.170
% of news words covered by GI	32,9%

Table 4. Evaluation of the general applicability of the translated GI on news items

7. Conclusion

The paper has presented the creation of a Slovene version of the General Inquirer which was translated with a mostly automated process. Several bilingual, aligned corpora and bilingual dictionaries have been used in order to make the translation as reliable as possible. The translation was done separately for words with single and multiple senses (as marked in the original GI). Single sense words are much easier to translate (multiple independent translations of the same word are compared; the word that is most frequent is selected as the translation). A problem however are words that are marked as single sense in the GI but have translations with multiple senses in Slovene. No mapping could be done automatically since single sense words have no sense description in the GI dictionary. This remains an open problem. For multi sense words each entry (in GI) is marked with additional description of the sense. These have been translated with the transformation to semantic networks and matching identical networks of word senses in both languages to find equality of meaning.

The translated dictionary was evaluated with regard to the correctness of the translation and the coverage of its entries on large scale Slovene corpora (140.000news items containing almost five million words). Both results show that the translated dictionary can be used for the sentimental analysis of Slovene texts. There are several things to consider when using a translated resource; foremost that the same words have different sentimental influence in different cultures. However the general valence (positive or negative) of words is language independent in most cases. Therefore translated resources can be used for the tasks of estimating the valence of the content.

Additionally to the translation of the dictionary we have performed statistical analysis of Slovene written language on a large scale which has resulted in the n-gram ($1<n<6$, unit is a single character) occurrence frequencies. This data can be used for language detection, enabling automated recognition if the analyzed content is in Slovene language.

8. References

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation, Valletta, MT, 2010, pages 2200-2204.

William B Cavnar and John M. Trenkle. N-gram-based text categorization. In Proc. SDAIR, pages. 161–175, 1994.

Kenneth R Beesley. Language identifier: A computer program for automatic natural-language identification of on-line text. In Languages at Crossroads: Proc.

Annual Conference of the American Translators Association, pages. 47–54, 1998.

Tomaž Erjavec and Darja Fišer. Building Slovene WordNet. In *Proceedings of the Fifth conference on International Language Resources and Evaluation LREC 06,* 2006, pages 1678-1683.

Tomaž Erjavec and Simon Krek. The JOS Morphosyntactically Tagged Corpus of Slovene. In *Proceedings of the 6th international Conference on Language Resources and Evaluation LREC08*. 2008, pages 322-326.

Tomaž Erjavec. MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh conference on International Language Resources and Evaluation,* Valletta, 2010, pages 1535-1538.

Tomaž Erjavec, Darja Fišer, Simon Krek and Nina

Ledinek. The JOS linguistically tagged corpus of Slovene. In *Proceedings of the Seventh conference on International Language Resources and Evaluation LREC10*, 2010 pages 1806-1809.

Vasileios Hatzivassiloglou and Kathleen R. McKeown.. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the ACL and Eighth Conference of the European Chapter of the ACL*, pp. 174-181, 1997.

George F. Luger. Artificial intelligence: Structures and strategies for complex problem solving , Addison-Wesley Longman, 2005.

George. A. Miller. Nouns in WordNet : A lexical inheritance system, *Int. J. of Lexicography*, 2(4), pp.245–264, 1990.

Alan G. Konheim. Cryptography: A Primer. JohnWiley & Sons. 46, 525, 1981.

Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources, In *Proceedings of the 9th Conference on Empirical Methods in Natural Language Processing* pp. 412–418, 2004.

Bo Pang, Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine-learning techniques, In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 79–86, 2002.

Ellen Riloff, Janyce Wiebe and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping, In *Proceeding of the 7th Conference on Natural Language Learning*, pp. 25-32, 2003.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions, In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 105–112, 2003.

Guang Qiu, Bing Liu, Jiajun Bu and Chun Chen, Expanding domain sentiment lexicon through double propagation, In *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (Morgan Kaufmann, San Francisco, pp. 1199–1204, 2009.

Philip J. Stone and Earl B. Hunt. A computer approach to content analysis: studies using the General Inquirer system. In *Proceedings of the May 21-23, 1963, spring joint computer conference*, pages 241-256. New York, NY, USA.

Peter D. Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp 417–434, 2002.

Cynthia Whissell. The dictionary of affect in language. *Emotion: Theory, Research, and Experience*, pp. 113–131, 1984.

Janyce M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2), pp. 233-287, 1994.

Tun T. Thet, J.C. Na and Christopher S.G. Khoo. Aspect-based sentiment of movie reviews on discussion boards. *Journal of Information Science*, 36(6), pp. 823-848, 2010.

Janyce M. Wiebe, Learning subjective adjectives from corpora, *Proceedings of the 17th National Conference on Artificial Intelligence*, pp. 735–740, 2000.

Root-Word Analysis of Turkish Emotional Language

Ozan Cakmak[1], Abe Kazemzadeh[2], Dogan Can[2], Serdar Yildirim[1], and Shrikanth Narayanan[3]

Department of Computer Engineering, Mustafa Kemal University[1]

Department of Computer Science, University of Southern California[2]

Department of Electrical Engineering, University of Southern California[3]

ozancakmak@mku.edu.tr, kazemzad@usc.edu, dogancan@usc.edu, serdar@mku.edu.tr,shri@sipi.usc.edu

Abstract

This paper describes a model for the perceived emotion of Turkish sentences based on the emotions associated with the constituent words. In our model, each emotion is mapped to a point in the continuous space defined by three emotional attributes: valence, activation, and dominance. We collected a large data set through two independent surveys: a word-level survey that prompted users with emotional words and asked them to assign each word a continuous emotional interval, and a sentence-level survey that prompted users with emotional sentences collected from 31 children's books and asked them to rate each sentence on a discrete emotional scale. The word-level survey was aimed at creating a core affective lexicon for Turkish. It is difficult to build a comprehensive affective lexicon for Turkish due to its very productive morphology that generates a very large vocabulary. We deal with the sparsity issues caused by the large word vocabulary by analyzing the emotional content of word roots. Our experimental results indicate that there is a strong correlation between the emotions attributed to Turkish word roots and the Turkish sentences.

Keywords: affective computing, emotion recognition, sentiment analysis, emotion analysis and annotation

1. Introduction

Automatically analyzing the emotional content of language has become increasingly important for applications that deal with natural language. For instance, the tasks of opinion mining and affective computing (Picard, 1997) are receiving a lot of attention in the fields of Natural Language Processing, Fuzzy Logic Systems such as an interval type-2 (Kazemzadeh et al., 2008), and Human Computer Interaction (Fragopanagos et al., 2005). Despite the progress of previous works (Jang and Shin,2010) in the field, there has been relatively less progress in non-English languages. The study of other languages within affective computing offers new technical and scientific challenges. We believe that our study opens new perspectives and brings about new methods that can increase the applicability of natural language affective computing to more diverse languages. In this study, we analyze Turkish(Katzner, 2002), which is an agglutinative language, which means that new words can be formed from existing words by a rich set of affixes (Oflazer, 1994).

The unique characteristics of Turkish present various challenges to current approaches to emotional analysis by natural language processing because the agglutinative word formation process create many unique words. In our study, we observed that there is a strong correlation between the emotions attributed to word roots, which are the core forms of words, and the emotion of sentences when negations, derivations, and inflections are accounted for. We measured correlation empirically from annotations of words and sentences in terms of valence, activation, and dominance (Russell and Mehrabian, 1977). A perennial challenge in affective computing research is the availability of suitable data resources. We have created a novel corpus of Turkish text from children's books richly annotated with affective information using crowd sourcing techniques. This corpus is large by the standards of other comparable emotional corpora and is one of the first emotionally labeled corpora for Turkish.

The reason why we chose this approach is that a single root might produce many different word forms in an agglutinative language like Turkish. Our hypothesis is that it is the emotion of the constituent word roots that determines and identifies the perceived emotion of sentences. However, such an analysis is not so simple because the effects of affixes like negation, which change the meaning of the roots, present theoretical contradictions to this general hypothesis. These affixes, which can potentially change the meaning of the root words, must be treated differently from the set of other affixes. Broadly, the affixation process can be seen in terms of phonological rules (e.g., vowel harmony, where vowel characteristics become assimilated in the neighboring vowels), derivational rules (e.g. grammatical recategorizations such as nominalization, which derives a noun from a verb), and inflectional rules (e.g. verb tense and noun pluralization).

2. Methodology

In our study, we analyze the data at both the word and the sentence level. Our sentence-level data comes from 31 children's books such as world classic novels, fairy tales, stories of heroism, romance, etc. Children's books were chosen to make up the corpus since these books contain a wide array of easily identifiable emotions. This corpus consists of 83,120 sentences. It contains 1,045,297 words, 110,695 of which are unique. The high number of unique words reflects the agglutinative nature of Turkish. The corpus is annotated at the sentence-level with one of the seven emotion categories (Angry, Happy, Sad, Disgusted,

Neutral, Surprised, and Fear) as well as valence, activation and dominance values, which are what we focus in this study. Valence measures whether the emotion is negative (unpleasant) or positive (pleasurable). Activation measures how strong the emotion is: dispassionate (calm) or passionate (excited). Dominance measures how assertive the emotion is: submissive (retreating) or dominant (aggressive). In our corpus, one point on the scale from 1-9 is used to represent these emotion characteristics. The corpus was distributed to 31 college students, who sequentially annotated the sentences with emotion category labels and valence, activation, and dominance values. To deal with the agglutinative word constructions, we extracted word roots with the Zemberek[1] Library, which is an open source, general purpose Natural Language Processing library for Turkish.

Let's take a look at example emotional words(Table 1-3):

Turkish	Öfkeli
English	Furious
Gloss	(root:öfke/fury) + (affix:-li/-ous, adjectival derivation)

Table 1: Example emotion word "furious"

Turkish	Hevesli
English	Zealous
Gloss	(root:heves/zeal) + (affix:-li/-ous, adjectival derivation)

Table 2: Example emotion word "zealous"

Turkish	Dertli
English	Sorrowful
Gloss	(root:dert/sorrow) + (affix:-li/-ful, adjectival derivation)

Table 3: Example emotion word "sorrowful"

After decomposing the words into root and affixes using the Zemberek Library(Akin et al., 2008), our corpus had 10,018 unique word roots. The word root can be seen as the basic component of a word's meaning after removing phonological, inflectional, and derivational effects. Our hypothesis is that the level of the root words is the best way to analyze Turkish sentences emotionally. However, stripping the words to their roots ignores critical derivations like negation. To measure the effects of these critical affixes, we performed two experiments: one, which removed these derivations, and another, which left them intact.

[1] http://code.google.com/p/zemberek/

English	**Turkish**
enthusiasm	Şevk
terrible	Berbat
courage	Cesaret
mad	Çılgın
tired	Yorgun
calm	Sakin
hopeful	Ümitli
interested	Ilgili
surprised	Şaşkın
boredom	Sıkıntı
sadness	Üzüntü
expectation	beklenti
worried	endişeli
lucky	Şanslı
happy	Mutlu
amusement	Eğlence
assidious	gayretli
confidence	Itimat
willing	Istekli
lucky	Şanslı
mercy	merhamet
patient	Sabırlı
love	Sevgi
joyful	sevinçli
admiration	Hayran
fear	Korku
frustration	Hüsran
arrogant	Kibirli
depression	depresyon
nervous	Sinirli
pleasure	memnuniyet
sympathy	sempati
proud	Gururlu
restful	huzurlu
excited	heyecanlı
heroism	kahramanlık
honorable	Onurlu

Table 4 : Some words from 197 Emotion Words

To measure the word-level emotion characteristics, we conducted a survey[2] of approximately 40 people who were presented with 197 emotion words (Table 4) and asked to rate these on valence, activation, and dominance scales. These words came from the EMO20Q Project(Kazemzadeh et al., 2011), which uses the emotion twenty questions game as a way to observe the human intuition about emotions. We translated 171 words from

[2] http://sail.usc.edu/~kazemzad/fuzzyEmotionEvaluation/turkish/turkish_experiment1.cgi

this project to Turkish and additionally added 26 synonyms. The emotional rating scales for this survey are different from the corpus annotation task in that two points are used for the scale, one to present the lower bound of a range of possible values and the other for the upper bound, which allows for measurement of intra-subject uncertainty. Also, the survey's scales ranged from 0 to 100. The survey consisted of four sessions per subject wherein each subject was presented with thirty-five words chosen randomly from the set of 197 words. This resulted in each of the 197 words being rated approximately 30 times. To compare the single-point scale of the sentence-level annotations to the double-point (upper and lower) scale of the word-level annotations, we converted the (upper-point, lower-point) representation into the (midpoint, radius) form.

Of the 197 emotion word roots from the survey, twenty-four did not occur in the corpus. As a result, the total count of word roots for the survey and the corpus is 173. In addition, in both the corpus and the survey, 99 emotion words were carefully chosen without possible derivational negations (the affixes -siz, sız, -suz and -süz), which can potentially change the emotion of word root. We separately analyze this subset and its complement.

Let's take a look at these examples (Table 5-6):

Turkish	ilgi -li
English	interested
Gloss	(root: ilgi/interest) + (affix: -li/-ed, adjectival derivation)
ANTONYM	
Turkish	ilgi-siz
English	un-interest-ed
Gloss	(root: ilgi/interest) + (affix: -siz/un-...-ed, negative adjectival derivation)

Table 5: Example emotion words "interested" and "uninterested"

Turkish	ümit-li
English	Hopeful
Gloss	(root: ümit/hope) + (affix: -li/-ful, adjectival derivation)
ANTONYM	
Turkish	ümit -siz
English	hope-less
Gloss	(root: ümit /hope) + (affix: -siz/-less, negative adjectival derivation)

Table 6: Example emotion words "hopeful" and "hopeless"

Although these words contain the same root, the derivational suffixes completely change the emotional connotation, in this case valence. To see the effects of these affixes, we performed correlation analysis both with and without these affixes.

3. Results

The 173 emotion word roots described above were identified in the corpus and the average sentence valence, activation, and dominance were calculated for each word root. Then, we compare, using correlation, these sentence-level averages with the word-level average valence, activation, and dominance values from the surveys. We found moderately high correlation between the word and the sentence-level valence (rho=0.55) and lower correlation for activation and dominance (rho=0.29 and rho=0.20, respectively). Then we repeated the correlation analysis on a subset of words having no negation present (99 words) and another subset having negation affixes (74 words).

Correlation	Valence
All words(173)	0.55
Words without negation and derivational affixes(99)	0.65
Words with negation and derivational affixes(74)	0.47

Table 7: Correlation Results for Valence.

Correlation	Activation
All words(173)	0.29
Words without negation and derivational affixes(99)	0.31
Words with negation and derivational affixes(74)	0.23

Table 8: Correlation Results for Activation.

Correlation	Dominance
All words(173)	0.20
Words without negation and derivational affixes(99)	0.24
Words with negation and derivational affixes(74)	0.10

Table 9: Correlation Results for Dominance.

We found that the subset without negation had a stronger correlation than the mixed set and the set containing negation affixes, and furthermore, that the set with negations had the lowest correlation values. This correlation of the averages of valence, activation and dominance values between the corpus and the survey indicates that perceived emotion of sentences is highly correlated with the chosen specific emotion words (Table

7-9). The stronger correlation in the valence dimension indicates that valence is the most strongly lexicalized emotional attribute.

4. Conclusion

In this paper, we verified that the emotions attributed to Turkish word roots are highly correlated with the emotion of Turkish sentences. We found that the emotional characteristics of sentences in terms of valence, activation, and dominance are significantly correlated with the emotional characteristics of the constituent words, when the words are decomposed into roots, and that moreover taking into account the exception of negation affixes makes this correlation stronger. This shows that negation affixes can significantly modify the emotion of words and sentences.

In our study, we measure the effects of this factor so that it can be taken into consideration in future studies. This approach of root analysis can be applied to various applications for extracting important emotions on the Internet, mobile phones or human computer interaction applications to make social networks for people who have similar opinions. Although English is not an agglutinative language, it also contains affixes that modify root words, so our results may be applied to non-agglutinative languages as well.

We plan to confirm the results of this paper by experiments on the survey and the corpus, which will be analyzed in more detail to consider negations, derivational affixes and inflectional suffixes. In addition to studying the relation of the word and sentence-level emotional scales, we also plan to examine the inter- and intra-subject variability. Inter-subject variability can be analyzed in terms of agreement between subjects and intra-subject variability can be seen in coherent behavior on repeated stimuli and by leveraging the upper and lower-points of the word-level surveys, which were designed for fuzzy logical analysis of emotional meaning.

Also, we plan to study the categorical labels of the sentence-level corpus. We plan to share this corpus, which is large by the standards of other comparable emotional corpora and one of the first emotionally labeled corpora for Turkish.

5. Acknowledgements

We are very grateful to all the annotators. This research was developed in the context of FLS (Fuzzy Logic System) for Turkish Language.

6. References

R. Picard. 1997. Affective Computing, MIT Press.

A. Kazemzadeh, S. Lee, and S. Narayanan. 2008. An interval type-2 fuzzy logic system to translate between emotion-related vocabularies, in Proceedings of Interspeech, (Brisbane, Australia).

N. Fragopanagos, J. G. Taylor. 2005. Emotion in human-computer interaction, Neural Networks, Volume 18, Issue 4, Pages 389-405.

H. Jang and H. Shin. 2010. Language-Specific Sentiment Analysis in Morphologically Rich Languages, Proceedings of Coling, (Beijing).

Kenneth Katzner. 2002. The Languages of the World, 3rd Ed., Routledge.

Oflazer Kemal. 1994. Two-level Description of Turkish Morphology, Literary and Linguistic Computing, vol. 9, No:2.

J. A. Russell and A. Mehrabian. 1977. Evidence for a three-factor theory of emotions. Journal of Research in Personality, vol. 11, pp. 273-294.

Ahmet Afsin Akin, Mehmet Dundar Akin.2007. Zemberek, an open source NLP framework for Turkish Languages.

A.Kazemdadeh, P.G. Georgiou, S. Lee, and S.Narayanan. 2011. Emotion questions: Toward a crowd-sourced theory of emotions, in Proceedings of AC'II.

A Chinese Corpus for Sentiment Analysis

Minlie Huang, Lei Fang, Xiaoyan Zhu

State Key Laboratory of Intelligent Technology and Systems,
Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology,
Tsinghua University, Beijing 100084, P.R. China
aihuang@tsinghua.edu.cn, jsfanglei@gmail.com, zxy-dcs@tsinghua.edu.cn

Abstract

Sentiment analysis and opinion mining has been a hot topic in the text mining and natural language processing communities. There have been a number of corpora in English or other western languages, either for sentiment classification, or for opinion extraction. However, to the best of our knowledge, few Chinese counterparts exist for these opinion mining tasks. In this paper, we introduce a Chinese corpus for opinion mining. The corpus contains two parts: a set of multi-domain sentences, with sentiment polarity annotated, and a set of multi-domain aspect-opinion pairs and corresponding polarities, which were obtained automatically from almost 5 million custom reviews. We present the corpus statistics, annotation guidelines, and discussions of how to use the corpus. We believe that such a corpus is potentially useful for sentence-level sentiment classification, aspect-level opinion extractions, opinion summarization, and so on.

1. Introduction

Many supervised or semi-supervised machine learning approaches for opinion mining require well-annotated corpora. In some sense, the availability of data resources has driven or limited the research of opinion mining and any other topics. Fortunately, there have been a number of efforts to providing such resources in English, such as (Hu and Liu, 2004)(Pang and Lee, 2004)(Pang and Lee, 2005)(Wiebe et al., 2005) (Wilson et al., 2005). However, to the best of our knowledge, there are only few Chinese counterparts, which has largely limited the research of Chinese opinion mining, cross-lingual or multi-lingual sentiment analysis with Chinese language.

We proposed a Chinese corpus for sentiment analysis in this paper. The corpus contains two parts: 1) A set of sentences, each of whose polarity has been annotated by three judges, according to four classes: Negative, Positive, Neutral, and Non-opinion - no opinion was expressed in a sentence. The sentences were obtained from several domains: digital products, finance news, entertainment news, and restaurant reviews. 2) A set of aspect-opinion pairs and corresponding polarities: the the pairs were extracted automatically from almost 5 million reviews and the polarity was predicted by our algorithm with a fairly high precision. The pairs contain three types of digital products: digital cameras, notebook computers, and cell-phones.

2. Sentence Polarity Annotation

We sampled about 1,000 sentences respectively from the digital, entertainment, and finance domain of the COAE 2011 corpus[1], and 3,463 sentences from the restaurant reviews of Dianping.com (See details in Table 1). For each sentence, we asked three annotators to judge the polarity. The polarity categories consists of the following classes:

- **Positive**: the sentence expresses positive opinions.
- **Negative**: the sentence expresses negative opinions.
- **Neutral**: the sentence expresses opinions but it's neither positive nor negative.
- **Non-opinion**: the sentence expresses some facts and contains no opinion at all.

2.1. Annotation Guidelines

Annotation tasks for opinion mining is absolutely challenging in that the annotation process is quite subjective since different people may have different cognitive understandings of opinions, emotions, and affections. In this case, an explicit, clear, and detailed guideline is indispensable. The guideline somehow decides the quality of annotation. Due to this concern, before the sentences are presented to annotators, we made an detailed guideline for the annotation process. For each polarity label, we give a clear definition and a list of typical examples. Three judges were asked to annotate the sentences for the initial pass. In the second pass, if there is inconsistency labeling, the annotators were asked to reach agreement by discussion.

Positive The sentence expressed a clear positive opinion toward some target. If the sentence contains also negative comments, the judge has to decide whether there is a remarkable bias to positive opinion. Here are two examples: (1) "就是稍贵了一点，但也算物有所值了(a little bit expensive, but it worths)" The first half is a negative comment, but the user's point is on the second half, so that this is a positive sentence. (2) "靠窗的景色更美，可惜没坐到靠窗的位子(The scene near window is more beautiful, but we did not have such a seat)" The second half describes a fact, while the user is more focused on the first half.

Negative Similar to **Positive**. A different case is like this example: "其它的菜色都很一般了(Other dishes are too soso)" where *soso* is a neutral word, but there is a quantifier modifying the word. For this example, we think it's a negative example.

Neutral First, the sentence expressed some opinion, but the polarity boundary between positive and negative is very vague. Some typical words such as "一般", "还可以", and

[1] http://ir-china.org.cn/coae2011.html

DataSet	Number of sentences				
	Negative	Positive	Neutral	Non-opinion	Total
restaurant reviews	432	1,581	70	1,380	3,463
digital products	56	362	3	583	1,012
finance news	42	70	3	841	957
entertainment news	29	193	2	775	1,004

Table 1: The statistics of sentence annotation.

Domain	#Positive Pairs	#Negative Pairs	Total Number
digital camera	3,210	2,272	5,482
notebook computer	2,872	2,427	5,299
cellphone	6,742	6,259	13,001

Table 2: The statistics of aspect-opinion pairs.

"还行" may signify this. However, if an neutral word is modified by a negation word or a quantifier (as seen in **Negative**), the judge should adjust the polarity accordingly.

Non-opinion The sentence did not express specific opinions on some target, and usually described some facts. Or if it were an opinionated sentence, we must figure it out that some target was said to be good or bad. Note, that we did not consider some factual sentence can express opinions. For example, "桌子上一把灰(there is much ash on the table)" might be viewed as negative opinion in some research.

3. Aspect-opinion Pair: Extraction and Polarity Prediction

3.1. Aspect Identification

Product aspect is discovered from a large number of reviews indexed in our review mining system - cReviewMiner [2]. cReviewMiner has indexed almost 10 million user reviews from major eCommerce websites in China, including 360buy.com, it168.com, zol.com, amazon.cn, and newegg.com. Noun phrases in reviews of digital camera, cellphone, and notebook computer were recognized, some simple filtering heuristics were applied, and the top frequent candidates were then presented to annotators. Once an aspect is identified, a k-means clustering algorithm is applied to cluster similar aspects. For example, Chinese terms such as "性价比","价位","价钱","价格","售价" refer to the same aspect "price". The central idea for this clustering approach is that aspects with similar contexts in the reviews should be grouped. The number of clusters is set to 20. Manual edits were performed on the automatically obtained clusters.

3.2. Aspect-opinion Pair Extraction

To extract aspect-opinion pairs, the central idea is that words which frequently appear in the left or right context of aspect terms might be an opinion word. The left and right contexts are respectively defined to the preceding and following 4 words of an aspect term. For instance, "高(high)" usually appears after the aspect term "价格(price)" with high frequency, hence "价格高" is extracted as an aspect-opinion pair. To simplify the problem, we limit opinion words to be adjective but adverbs will be considered in extracting aspect-opinion pairs.

First of all, the review text was processed with Chinese word segmentation and part of speech tagging. Then, we count the occurrences of adjectives and adverbs adjacent to an aspect term. Some frequent patterns were automatically discovered from the data, such as "adjective+adjective", "adverb+adjective" and "adverb+verb". These patterns are then used to merge the frequencies of different instances that belong to the same aspect-opinion pair. For example, "价格非常高(price is very high)", "价格比较高(price is comparatively high)", and "价格太高(price is too high)" are all belonging to the aspect-opinion pair <价格(price),高(high)>. Negation words are also considered in this process. For example, "价格不高(price is not high)" is also merged into the previous examples. Finally, the pairs with high frequency are extracted as resultant aspect-opinion pairs.

3.3. Pair Polarity Prediction

Polarity prediction of aspect-opinion pairs benefits from the large number of user reviews with polarity labels. In our data, a user who wrote a review has already assigned positive and negative labels for addressing the advantages and disadvantages respectively. This was actually required by most eCommerce websites. Our assumption is that if an pair appears relatively more frequently in positive reviews, its polarity is positive; and if it appears relatively more frequently in negative reviews, its polarity is negative. Though many users sometimes put negative comments in positive labeled reviews (or vice versa), we found this method is very accurate, with about 98% accuracy. Some examples for each domain are shown in Table 3.

3.4. Discussion

The dataset might be useful for context-aware opinion mining. Different from polarity lexicon, the pairs are domain-independent, and the polarity has attached to some specific aspect. For example, word "高（high）" has positive polarity in most lexicons, however, in our dataset, "温度高(high temperature)" is a negative term, while "性价比高

[2] http://166.111.138.18/cReviewMiner/ or http://www.qanswers.net:1880/cReviewMiner/

Domain	Phrase (**Aspect_word** Opinion_word)	Polarity
digital camera	**光圈** 很棒(great aperture)	positive
digital camera	**光圈** 不够大(aperture is not large enough)	negative
digital camera	**按键** 太敏感(too sensitive keys)	negative
digital camera	**按钮** 还不错(button is good)	positive
digital camera	**触摸屏** 不灵敏(touch screen is not sensitive)	negative
digital camera	**屏幕** 清晰(screen is very clear)	positive
cellphone	**信号** 不稳(signal is not stable)	negative
cellphone	**通话** 比较清晰(speech signal is clear)	positive
cellphone	**续航** 比较短(battery life is comparatively short)	negative
cellphone	**耗电** 太猛(battery consume is too quick)	negative
cellphone	**输入法** 非常麻利(IME is very easy to use)	positive
cellphone	**手写** 不好(handwriting is not good)	negative
notebook	**电池**有点小重(battery is a little bit heavy)	negative
notebook	**续航** 出色(battery life is excellent)	positive
notebook	**硬盘**很安静(disk is quiet)	positive
notebook	**硬盘** 较慢(disk is slow)	negative
notebook	**显卡** 比较强劲(graphic card is powerful)	positive
notebook	**集显** 差(poor integrated graphics)	negative

Table 3: The examples of aspect-opinion pairs.

（high value-to-price ratio）" is a positive term. In other words, the polarity of a term depends not only the domain of interest, but also the aspect it was attached.

Therefore, the dataset may be used as features in sentiment classification to involve context factors, for example, to improve bag-of-unigram models. Further, the aspect-opinion pairs may help do aspect summarization, aspect identification, and aspect ranking.

4. Conclusion and Discussion

We presented a Chinese corpus for sentiment analysis. The corpus contains two parts: a set of multi-domain sentences with annotated polarity, and a set of aspect-opinion pairs obtained from three types of digital products including digital camera, notebook computer, and cellphone. We described the annotation guideline of labeling these sentences, and the extraction process of obtaining aspect-opinion pairs. Such a dataset would be useful in sentiment classification, cross-domain transfer learning, or context-aware opinion mining, as discussed.

The dataset will also be supportive in cReviewMiner. Ongoing research includes cross-domain sentiment classification and context-aware opinion mining.

5. Acknowledgements

This paper was supported by Chinese 973 project under No. 2012CB316301 and National Chinese Science Foundation project with No. 60973104.

6. References

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 168–177, New York, NY, USA. ACM.

Bo Pang and Lillian Lee. 2004. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 115–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, pages 165–210.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT/EMNLP'05*, pages –1–1.

Laughter Annotations in Conversational Speech Corpora – Possibilities and Limitations for Phonetic Analysis

Khiet P. Truong[1] & Jürgen Trouvain[2]

[1] University of Twente, The Netherlands & [2] Saarland University, Germany

E-mail: [1]k.p.truong [at] utwente.nl & [2]trouvain [at] coli.uni-saarland.de

Abstract

Existing laughter annotations provided with several publicly available conversational speech corpora (both multiparty and dyadic conversations) were investigated and compared. We discuss the possibilities and limitations of these rather coarse and shallow laughter annotations. There are definition issues to be considered with respect to speech-laughs and the segmentation of laughs: what constitutes one laugh, and when does a laugh start and end? Despite these issues, some durational and voicing analyses can be performed. We found for all corpora considered that overlapping laughs are longer in duration and are generally more voiced than non-overlapping laughs. For a finer-grained acoustic analysis, we find that a manual re-labeling of the laughs adhering to a more standardized laughter annotations protocol would be optimal.

1. Introduction

Laughter is a non-verbal phonetic activity that usually occurs in conversational interaction with an interlocutor. In contrast to this we can state that most studies on the acoustics of laughter were *not* based on conversational settings but settings in which actors produce pre-selected laughter categories (Habermann 1955; Szameitat et al. 2009) or in which subjects watch funny video clips, either alone (Urbain et al. 2010) or with another person (Bachorowksi et al. 2001).

One important social feature of laughter *in conversations* is that it frequently is a joint action of two persons. Subsequently, laughs of interlocutors often overlap with laughs of the other. Since we are interested in studying phonetic and social aspects of laughter in conversation, of which overlapping laughter represents an important aspect, the first step to be taken is to look for laughter in conversational speech corpora.

Most studies focusing on laughter in conversations are based on rather restricted amounts of data either investigating actors in movies (Pompino-Marschall et al. 2007), focusing on interviews in mass media (O'Connol & Kowal 2004), eliciting experimental data, e.g. on male-female encounters (Grammer & Eibl-Eibesfeldt 1990) or on mother-child interaction (Nwokah et al. 1999), analysing a small corpus of acted dialogues recorded in a push-to-talk mode (Trouvain 2000), or performing qualitative studies of conversational analysis with only a few examples (e.g. Jefferson 1985).

Studies with larger data sets are often not publicly available, such as the natural dyadic conversations used in Vettin & Todt (2004). And sometimes, the conversations are recorded in a language unknown to the researchers that can be rather inconvenient, such as the recordings in Japanese used in Campbell (2007) where strangers have repeated telephone calls with each other.

There are a number of large conversational speech corpora publicly available containing laughter but usually, the developers of these databases did not record these with the aim to study laughter or other paralinguistic phenomena. Therefore, often only coarse and shallow annotation of laughter is available because only little attention was given for how to label laughter. Consequently, we cannot expect to find a standard labelling of laughter across multiple corpora.

In this study, we explore laughter annotations in different speech corpora and show how these can be used for phonetic analysis. The aims of this study are three-fold: 1) to compare and select different corpora suitable for phonetic laughter analysis, 2) to identify difficulties in laughter labelling, 3) to show how shallow laughter annotations can be used to explore durational and voicing aspects of overlapping laughter in conversation.

2. Conversational speech corpora

Prerequisites of conversational speech corpora ideally comprise: 1) separated channels for each speaker, 2) searchability of annotated laugh events in the transcription document, 3) time alignment of transcription and audio file with time stamps for the beginning and the end of the laugh event, 4) publicly available.

Not all corpora meet the mentioned criteria such as the separation of the recording channels. An example for a corpus with one channel for all speakers is the Santa Barbara Corpus of Spoken American English (SBC). Another example is the Buckeye corpus (Pitt et al. 2007)

for which only the data of the interviewed person is available but not the data of the interviewer as the interlocutor. The disadvantage of having only one channel is that during overlapping signals like cross-talk or overlapping laughs it is not clear which part of the signal stems from which speaker. However, for a fine-grained acoustic and temporal analysis this intertwining of both speakers can be very important as illustrated in Fig. 1 (taken from the Diapix Lucid corpus (Baker & Hazan 2011)).

Corpora can differ very much with respect to the annotation of laughter. For two larger Dutch conversational speech corpora, CGN (Oostdijk 2000) and IFADV (van Son et al. 2008) laughter was annotated with a label that also comprised other types of non-verbal vocalizations, e.g. coughs.

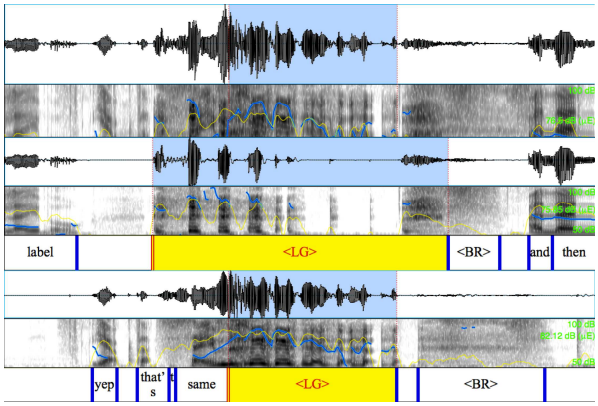

Figure 1: Example of an overlapping laugh (waveform and spectrogram. Top: mixed signal with masked information of speaker identity. Middle: signal of speaker A. Bottom: signal of speaker B.

Even if laughter was somehow annotated in the transcription files, the laughter annotations sometimes cannot readily be used for signal analysis because of missing ending times of laugh events (e.g. Lindenstraße corpus IPDS 2006).

In selecting suitable speech corpora, we restricted ourselves to the English language. However, the considered corpora do not represent an exhausted list because availability of data depends e.g. on financial aspects. We selected 4 corpora that met our prerequisites: the AMI meeting corpus (Carletta et al. 2007), the ICSI meeting corpus (Janin et al. 2003), the HCRC Map Task Corpus (Anderson et al. 1991), and the Diapix Lucid corpus (Baker & Hazan 2011), see also Table 1. The first two corpora contain multi-party meeting recordings and the latter two consist of task-based dyadic conversations. The main reason for considering 4 different corpora that we wanted to test how general our findings are.

3. Laughter annotations

We manually inspected some of the laughter annotations in the four mentioned corpora and encountered a number of problems in the annotations.

3.1 Definition problems

1. Are speech-laughs considered as a sub-type of laughs?
Sometimes speech-laughs are ignored and sometimes they are inconsistently labeled.

2. What is the definition of one laugh?
Sometimes the annotated laugh is in reality composed of two or more laughs, and vice versa, two annotated laughs are in reality one laugh. It also happens that the annotated laugh is only partially a laugh or sometimes it is unclear whether it was a laugh or not.

3. When does the laughter event start and end?
Sometimes the annotated laughs show incorrect time stamps for beginning and/or end.

3.2 Other problems

1. Are all audible laughs annotated?
Sometimes laughs in the audio file were missing in the annotation.

2. Are there any technical errors?
Sometimes there were annotated laughs with negative durations, or no timestamps at all.

Exploiting the information about laughter needs clear labelling criteria and a consistent application of these criteria. It seems to be that human annotation is better than annotations obtained by a machine (i.e. automatic forced alignment). In any of the corpora inspected we would consider a re-annotation as necessary to obtain more homogeneous laughter annotations across corpora that in turn will lead to more consistent and reliable research results.

4. Laughter analysis

Despite the listed drawbacks the existing corpora can be used as they are – but always with the restriction that we are not considering completely correct data.

4.1 Data used

The laughs used in the analysis were automatically extracted based on the transcriptions available from the four corpora under inspection. Speech-laughs were sometimes annotated in the corpora, e.g., ICSI meeting corpus (Janin et al. 2003) and Diapix Lucid corpus (Baker & Hazan 2011), but these were discarded in our analysis to make the data comparable to the HCRC Map Task corpus (Anderson et al. 1991) and the AMI meeting corpus (Carletta 2007). The transcribed laughs were most of the times treated as words with starting and ending times. However, a subpart of the annotated laughs was discarded due to missing time stamps, missing transcriptions or other technical issues. Since we are investigating overlapping laughs, only those laughs that have a start and end time were included in our analysis. Table 1 gives a short description of the corpora and laugh data used.

Table 1: Descriptive features of inspected corpora.

	no. of annotated laughs	no. of used laughs	no of speakers	no. of convers.	no. of speakers per convers.	mean duration of convers. (in mins)	task	visual contact	relationship between speakers
AMI	16477	8803	679	171	3-4	35.1 (13.5)	acted meeting	yes	mostly strangers
ICSI	12574	8388	494	75	3-11	55.0 (15.9)	real meeting	yes	colleagues
HCRC	1002	966	250	125	2	6.8 (3.1)	giving route on a map	yes/no	friends + strangers
DiaPix	582	575	114	57	2	7.7 (2.3)	spot-the-difference	no	friends

4.2 Frequency of occurrence

Fig. 2 reveals that overlapping laughs represent a substantial part of all laughs in all corpora ranging from 35% to 63% of all annotated laughs. Only the ICSI corpus shows more overlapping than non-overlapping laughs. This can be easily explained by the fact that in the ICSI corpus there are many more persons present and thus increasing the probability that two speakers will overlap with their laughs. Additionally a 'contagious effect' could be at work for laughter as was already shown by Laskowski & Burger (2007b).

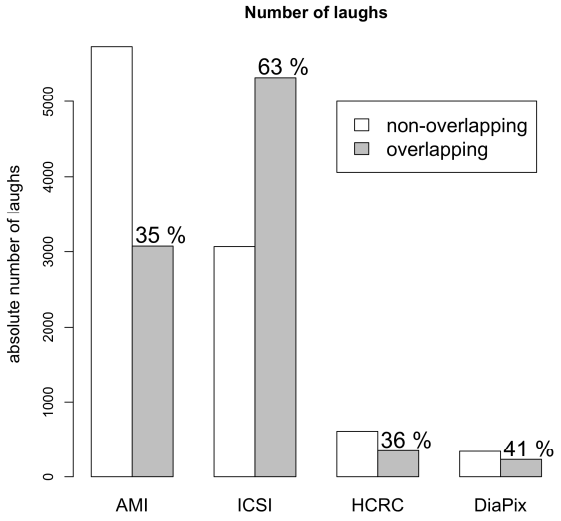

Figure 2: Frequency of occurrence of non-overlapping and overlapping laughs for each corpus. Percentages indicate the relative number of overlapping laughs.

4.3 Duration

The descriptive statistics illustrated in Table 2 and Fig.3 clearly show that overlapping laughs are longer than non-overlapping laughs. T-tests reveal that for each corpus these durational differences reach statistical significance at $p<0.01$. Interestingly, the multi-party meetings show higher durations in average, at least for overlapping laughs. The ICSI corpus differs again compared to the others by showing longer mean durations for overlapping as well as for non-overlapping laughs.

Table 2: Mean duration and standard deviation in seconds of all laughs (left), non-overlapping laughs (NO) and overlapping laughs (OL) pooled over the inspected corpora.

	all		NO		OL	
	mean	sd	mean	sd	mean	sd
AMI	1.042	1.184	0.775	0.842	1.541	1.521
ICSI	1.661	1.298	1.195	0.753	1.929	1.460
HCRC	0.838	0.652	0.715	0.524	1.052	0.784
DiaPix	0.899	0.689	0.755	0.495	1.107	0.860

4.4 Voiced vs. unvoiced laughter

Laughter is sometimes classified in voiced vs. unvoiced forms (e.g. Grammer & Eibl-Eibesfeldt 1990, or Bachorowski et al. 2001). For our analysis we define those laughs as unvoiced that show no voiced frame at all (as obtained from a pitch analysis with a window length of 40 ms and time step of 20 ms). The rest of the laughs are defined as "voiced" even if the number of voiced frames can be relatively low (in contrast to Laskowski & Burger (2007a) who did a manual classification of voicedness leading to a higher number of unvoiced laughs for the ICSI corpus).

In Fig. 4, we can observe a positive correlation between the level of voicing and duration (similar to Laskowski & Burger 2007a). There are hardly any unvoiced laughs longer than 1.6 sec and most unvoiced laughs are shorter than 800 ms.

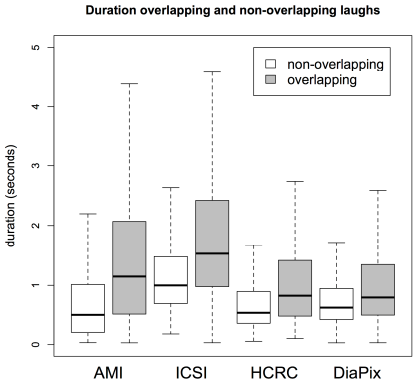

Figure 3: Boxplots of the duration in seconds of non-overlapping and overlapping laughs in the four inspected corpora. Outliers were computed but not shown for illustrative reasons. Whiskers indicate 1.5*inter-quartile range of the data.

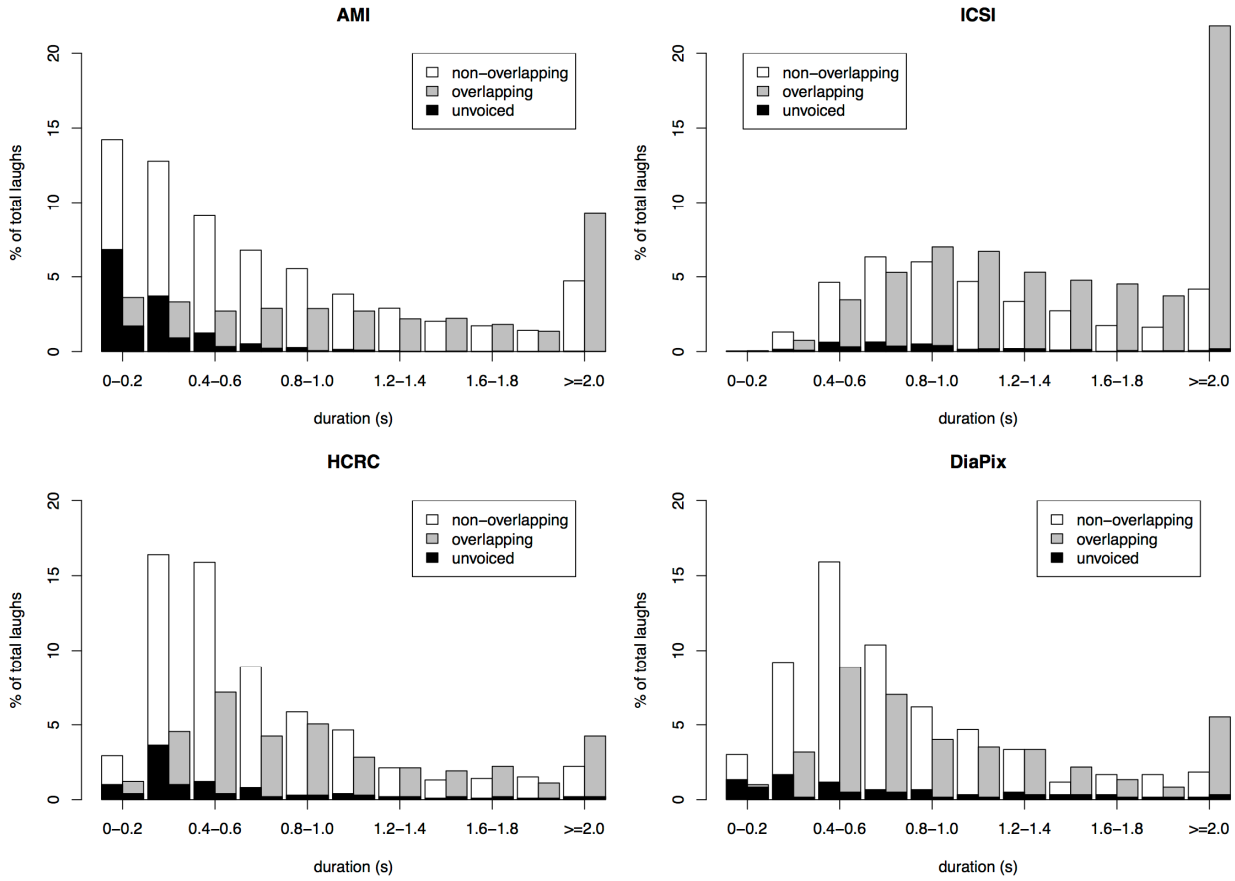

Figure 4: Histograms (for each corpus) of non-overlapping vs. overlapping laughs distinguishing unvoiced and "voiced" laughs in bins of 200 ms.

Fig. 4 also shows for all four corpora that the longer the laugh the higher the probability that the interlocutor joins in, resulting in an overlapping laugh. This effect is clearest for the ICSI meeting corpus where up to 11 conversational partners were present. For this corpus there are also the fewest unvoiced laughs counted in relation to the total number of laughs.

5. Concluding Remarks

In comparing conversational speech corpora we have found differences in the duration and numbers of overlapping laughs between corpora, particularly between multi-party conversations and dialogues. In general we could observe the tendency that overlapping laughs are more likely to be longer than non-overlapping ones; we hypothesize that this has to do with the social function of laughing together. In addition we saw that among the shorter laughs there was a relatively high proportion of unvoiced laughs.

The "noise" of the laughter annotations could have influenced results but the observations are made in multiple corpora giving strong evidence for our conclusions. However, we still consider a manual re-labelling of the laughter annotations as optimal for further more fine-grained acoustic analyses. Future research should include looking at acoustic characteristics of various kinds of laughter (overlapping vs. non-overlapping, voiced vs. unvoiced, speech-laughs), in addition to duration.

Acknowledgements

This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet) and the UT Aspasia Fund.

6. References

Anderson, A.H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weintert, R. (1991). The HCRC Map Task Corpus. *Language and Speech* 34(4), pp. 351-366.

Bachorowski, J.-A., Smoski, M.J., Owren, M.J. (2001). The acoustic features of human laughter. *Journal of the Acoustical Society of America* 111(3), pp. 1582-1597.

Baker, R, Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3), pp. 761-770.

Campbell, N. (2007). Whom we laugh affects how we laugh. *Proc. Workshop on "The Phonetics of Laughter"*, Saarbrücken, pp. 61-65.

Carletta, J.C. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and*

Evaluation 41(2), pp. 181-190.

Grammer, K., Eibl-Eibesfeldt, I. (1990). The ritualisation of laughter. In: Koch, W.A. (Hrsg.) *Natürlichkeit der Sprache und der Kultur: acta colloquii*. (Bochumer Beiträge zur Semiotik; 18) Bochum: Brockmeyer, pp. 192-214.

Habermann, G. (1955). *Physiologie und Phonetik des lauthaften Lachens*. Leipzig: J. A. Barth.

IPDS (2006). *Video Task Scenario: Lindenstraße* – The Kiel Corpus of Spontaneous Speech, Volume 4, DVD, Institut für Phonetik und Digitale Sprachsignalverarbeitung Universität Kiel.

Janin, A., Baron, D., Edwards, D., Ellis, D., Gelbart, D., Morgan, N. (2003). The ICSI meeting corpus. *Proceedings of ICASSP*, pp. 364-367.

Jefferson, G. (1985). An exercise in the transcription and analysis of laughter. In T. Van Dijk (Ed.) *Handbook of discourse analysis*, Vol. 3: *Discourse and dialogue* (pp.25-34).London, UK: Academ. Pr.

Laskowski, K., Burger, S. (2007a). On the correlation between perceptual and contextual aspects of laughter in meetings. *Proc. Workshop on "The Phonetics of Laughter"*, Saarbrücken, pp. 55-60.

Laskowski, K., Burger, S. (2007b). Analysis of the occurrence of laughter in meetings. Proceedings of Interspeech, Antwerp, pp. 1258-1261.

Nwokah, E.E., Hsu, H.-C., Davies, P. & Fogel, A. (1999). The integration of laughter and speech in vocal communication: a dynamic systems perspective. *Journal of Speech, Language and Hearing Research* 42, pp. 880-894.

O'Connell, D. C., Kowal, S. (2004). Hillary Clinton's laughter in media interviews. *Pragmatics* 14(4), pp. 463-478.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. *Proc. LREC*, pp. 887-894.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E. and Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu] Columbus, OH: Department of Psychology, Ohio State University (Distributor).

Pompino-Marschall, B., Kowal, S., O'Connell, D. (2007). Some Phonetic Notes on Emotion: Laughter, Interjections, and Weeping. *Proc. Workshop on "The Phonetics of Laughter"*, Saarbrücken, pp. 41-42.

Santa Barbara Corpus. http://www.linguistics.ucsb.edu/research/sbcorpus.html retrieved 15 Feb 2012

Szameitat, D.P., Alter. K., Szameitat. A.J., Dietrich. S., Wildgruber. D., Sterr. A., Darwin C.J. (2009). Acoustic profiles of distinct emotional expression in laughter, *Journal of the Acoustical Society of America* 126, pp. 354-366.

Trouvain, J. (2001). Phonetic aspects of 'speech-laughs'. Proceedings of the *2nd Conference on Orality & Gestuality (ORAGE)*, Aix-en-Provence, pp. 634-639.

Urbain, J., Bevacqua, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., Tilmanne J., Wagner, J. (2010). The AVLaughter-Cycle Database. *Proc. LREC*, Malta, pp. 2996-3001.

Van Son, R., Wesseling, W., Sanders, E., van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. *Proc. LREC*, pp. 501-508.

Vettin, J., Todt, D. (2004). Laughter in conversation: features of occurrence and acoustic structure. *Journal of Nonverbal Behaviour* 28(2), pp. 93-115.

Finding out the audio and visual features that influence the perception of laughter intensity and differ in inhalation and exhalation phases

Radosław Niewiadomski[1], Jérôme Urbain[2], Catherine Pelachaud[1], Thierry Dutoit[2]

[1]LTCI-CNRS Telecom ParisTech
37 rue Dareau, 75014 Paris, France
[2]Université de Mons, Faculté Polytechnique, TCTS Lab
20, Place du Parc - 7000 Mons, Belgium

{niewiado,catherine.pelachaud}@telecom-paristech.fr
{jerome.urbain,thierry.dutoit}@umons.ac.be

Abstract

This paper presents the results of the analysis of laughter expressive behavior. First we present the intensity annotation study of an audiovisual corpus of spontaneous laughter. In the second part of the paper we present the analysis of audio and visual cues that influence the perception of laughter intensity, as well as the study of audio and visual features that differ in laughter inhalation and exhalation phases.

Keywords: laughter, audiovisual synthesis, intensity

1. Introduction

Several research works on social signals were recently undertaken with possible applications in latest HCI technologies such as virtual agents. Laughter is one such signal. It occurs frequently in human-human interaction, and may have many functions and meanings, such as being the expression of some emotional states, as well as having a social function (Adelsward, 1989). Surprisingly enough, virtual agents - software created to be able to maintain natural multimodal verbal and nonverbal interaction with humans - are still not able to laugh. Knowledge about the expressive patterns of laughter is still limited. Within the long term aim of building a laughing virtual agent, this paper presents the results of our ongoing work on the analysis of laughter expressive behavior. We report on the annotation of an audiovisual corpus of spontaneous laughter, on a study of audio and visual cues that influence the perception of laughter intensity, as well as on a study of audio and visual features that differ in laughter inhalation and exhalation phases.

This paper is structured as follows. In next Section we explain the motivation of this research. Section 3. is dedicated to the description of the intensity annotation protocol. Then, in Section 4. we present the data analysis that we realized so far whereas in Section 5. we present the detailed results. Finally we conclude the paper in Section 6.

2. Motivation for this work

Multimodal laughter synthesis is a complex task. In laughter, the body movements and the tight synchronization between audio and visual signals of the expression is crucial. Laughter is a highly multimodal expression composed of very quick rhythmic shoulders and torso movements, visible inhalation, several facial expressions which are often accompanied with some rhythmic as well as communicative gestures (Ruch and Ekman, 2001). This makes its synthesis particularly challenging. Recent studies on laughter suggest that there exist different types of laughter that can have different expressive patterns (Huber et al., 2009). Consequently, even a small incongruence in laughter synthesis may influence its perception. Particular attention has to be put on the synchronization between modalities which seems to be the key factor of successful laughter synthesis. Thus we need to study first the synchronization between modalities in human laugh acts.

Even less is known about which audio and visual cues influence the perception of laughter intensity. Differently to many other expressive behaviors studied so far, laughter is a highly multimodal expression. We expect that for laughter the perceived intensity should be a global evaluation that takes into consideration all single monomodal signals. Thus measuring only audio loudness or only mouth openness is not enough to define laughter intensity. Obviously the knowledge about these audio and/or visual cues that influence laughter intensity perception is indispensable in realistic laughter synthesis. In order to properly model laughter in virtual agents, we first need to find the factors that influence the perception of the intensity of human laughs.

In this paper we describe the results of studies aiming to better understand the expressive patterns of human laughter. We mainly focus on the intensity of laughter. For the purpose of this study we used the AudioVisualLaughterCycle (AVLC) corpus (Urbain et al., 2010) that contains about 1000 spontaneous audio-visual laughter episodes with no overlapping speech. The episodes were recorded with the participation of 24 subjects. Each subject was recorded watching a 10-minutes comedy video. Smart Sensor Integration (Wagner et al., 2009) was used to acquire the signals and manually annotate (and segment) the laughter episodes. The number of laughter episodes for a subject varies from 4 to 82. Each episode was captured with one of two motion capture systems (Optitrack or Zigntrack) and synchronized with the corresponding audiovisual sample.

Each segmented laugh was also phonetically annotated (Urbain and Dutoit, 2011). Two annotation tracks were used: one to indicate the airflow direction (inhaling or exhaling), the other for the actual phonetic transcription.

3. Intensity annotation

We conducted an annotation study of laughter intensity of the AVLC database. The annotation was realized through a web application. This application is composed of a set of web pages; each of them displays one AVLC episode. Participants to this study were asked to give an overall score of their perceived intensity of the episode using a Likert scale from 1 (low intensity) to 5 (high intensity). Each laugh episode of AVLC was evaluated globally with only one score. There was no obligation to annotate all the available examples (352 episodes). There was no time limitation for the annotation task. Participants could see each sample several times. Once they had evaluated an episode and gone to another one they could not change their previous score. The episodes were displayed in random order. The whole set of episodes was divided into subsets, each of them containing the episodes corresponding to 4 subjects.

For the moment, 2 subsets of the whole database (i.e. 352 out of 995 episodes corresponding to 8 subjects) have been annotated by 15 naive participants mainly from France and Belgium, aged 24-40. Each episode has been annotated by at least 3 and at most 6 coders. Overall agreement between coders was fair: Krippendorff's alpha (Krippendorff, 2012) was .66.

In total we collected 1661 answers. The distribution of the intensity scores in the part of database annotated so far is not uniform. Most of the episodes were evaluated as low intense (see Figures 1 and 2). In more details, the lowest intensity value was used 536 times, score 2 was used 512 times, 3 - 352, 4 - 222, and the maximal score has only been given 39 times.

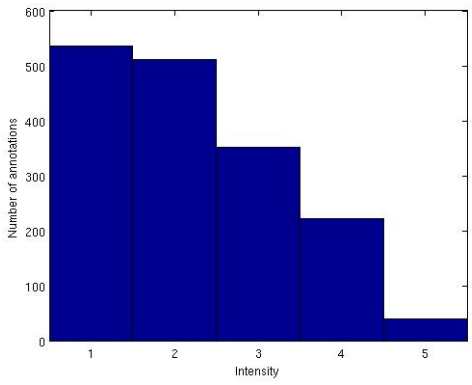

Figure 1: Laughs intensity annotations histogram

4. Data analysis

In this work we focused on two research questions:

- T1) the relation between the perceived intensity and certain audio and/or visual features,

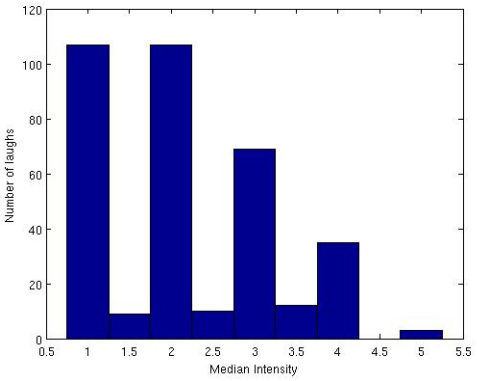

Figure 2: Number of episodes for each degree of intensity

- T2) the relation between the respiration phases and certain audio and/or visual features.

Task T1. The first task relies on the annotation of perceived intensity of laughter (see Section 3.). We aim to discover audio and visual features that correlate with the different degrees of intensity. For each episode we extract several distances between markers that correspond to some action units (Ekman and Friesen, 1978) as well as low-level acoustic descriptors. We are particularly interested in the audio and visual features that can be associated with intense laughs (such as maximum mouth opening).

Task T2. The second task relies on the annotation of respiration phases in the laughter episodes. Respiration has an important role in the multimodal laughter expression. We expect that information about respiration is crucial to achieve believable audiovisual laughter synthesis: indeed, humans can naturally distinguish these respiration phases when listening or watching to a laugh. The audiovisual signals of the two respiration phases must thus present different patterns. If so, this information can be later used to drive the audio and visual synthesis modules with a common respiration input signal, ensuring the synchronization between the characteristic audio and visual patterns of the two respiration phases. To verify this hypothesis, we analyze the relation between the respiration phases and our audio and visual features and we check if these features take different values in the two respiration phases.

The extracted characteristics are 12 distances corresponding to some facial actions and 58 acoustic low-level descriptors:

- Facial actions are characterized by distances between the markers in the motion capture data. The computed distances correspond to jaw movement (D1), lip height (D2), lip width (D3), cheek raising (D4-5), upper lip protrusion (D6), lower lip protrusion (D7), lip corner movement (D8-9), frown (D10-12). The measurements D4-D5 and D8-D9 roughly correspond to action units considered to be specific for the facial expression of hilarious laughter, namely cheek raising - AU 6 and smile (lip corner up) - AU 12. The remaining measurements correspond to the action units which occurrence in certain laughs is optional or it is still discussed

(Drack et al., 2009) such as AU4 (frowning) or AU 25 (mouth opening) and AU 26 (dropping the jaw). All these characteristics are computed at 25 FPS.

- Acoustic low-level descriptors can be divided into 3 categories: spectral low-level descriptors, measures of the noise level and prosody-related low-level descriptors. Spectral low-level descriptors are 13 MFCCs (as well as their first and second order derivatives), spectral centroid, spectral spread, spectral decrease, spectral flux and spectral variation. Measures of noise are obtained with Harmonic to Noise ratios (HNR, 4 values corresponding to the frequency bands 250-500Hz, 500-1000Hz, 1000-2000Hz and 2000-4000Hz), spectral flatness (4 values also), cepstral peak prominence, chirp group delay and zero crossing rate. Finally, prosody-related low-level descriptors include measures of energy and fundamental frequency. Further details about these low-level descriptors can be found in (Drugman et al., 2011; Peeters, 2003). All these acoustic low-level descriptors were extracted from the 16kHz audio signals, using windows of 512 samples (32ms) shifted by 160 samples (10ms).

For each considered segment (full episode and respiration phase respectively for Task T1 and T2), the frame by frame low-level descriptors (in variable number, depending on the duration of the segment) are mapped to a fixed-length feature vector with the help of the following functionals: minimum over the segment, max, range, mean, standard deviation, skewness, kurtosis, percentage of time spent in the upper quartile ($\%25$), zero-crossing rate (ZCR). Since we had 12 facial distances and 58 acoustic low-level descriptors, we obtain a feature vector of 630 audiovisual features per segment, plus the duration of the segment.

5. Results

We present the results based on the subset of the AVLC corpus for which we have sufficient intensity annotations (see Section 3.). Two subjects had to be removed from the current study due to erroneous motion capture data. Consequently, we had 1336 intensity annotations for the remaining 249 laughs (from 6 subjects).

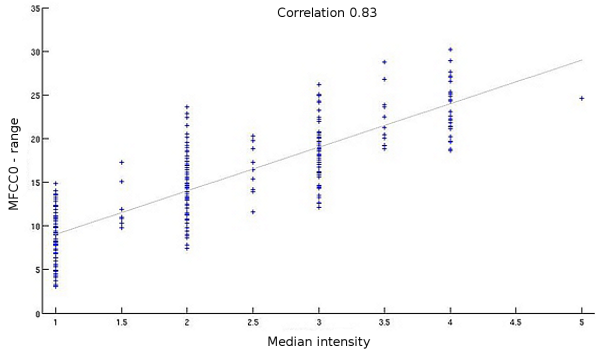

Figure 3: Correlation between median intensity and MFCC0 range

5.1. Intensity and audio visual features

In task 1 we studied the relation between the perceived intensity and several audio and visual features. Concerning the audio features we found strong correlations between several features and the median intensity annotated for each laugh. Spectral features provide the strongest correlations, as well as energy: MFCC0 presents a correlation coefficient (ρ) with the laughter intensity above .8, while loudness is slightly behind. Figures 3 and 4 show the best correlations with the annotated intensity, obtained with MFCC0 range and MFCC2 range, respectively. The detailed data for the 10 best audio descriptors and pitch are presented in Table 1. We can see that the "range" functional is yielding the best correlations for all these lew-level descriptors. Energy descriptors ($MFCC0$, $\Delta MFCC0$, $\Delta\Delta MFCC0$ and Loudness) are the most correlated with laughter intensity, followed by descriptors of the spectral shape (spectral flatness and MFCCs). Pitch, extracted through the ESPS method available in Wavesurfer (Sjölander and Beskow, 2011), is slightly below.

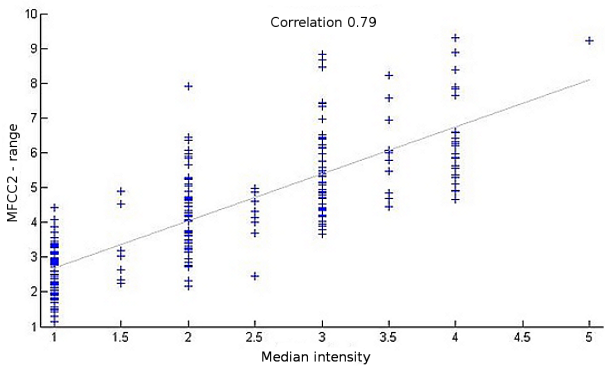

Figure 4: Correlation between median intensity and MFCC2 range

Visual features give slightly lower correlation coefficients. The strongest correlation was observed for the maximum jaw (Figure 5) and lip openings, i.e. the distances D1 and D2, with the "max" functional computed on the whole episode ($\rho = .68$ and .65, respectively).

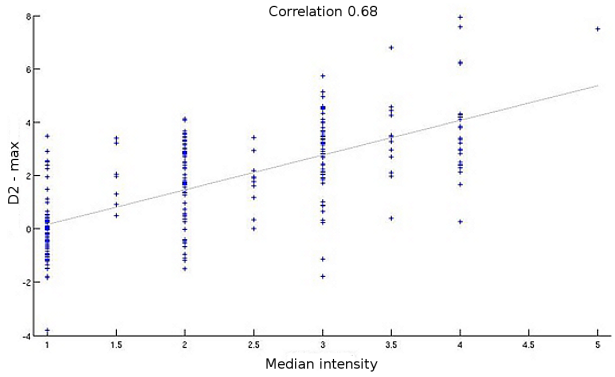

Figure 5: Correlation between median intensity and jaw opening

Table 1: Correlation between laughter median intensity and the 10 best acoustic descriptors (+ pitch)

	SF 1-2kHz	SF 2-4kHz	MFCC0	MFCC2	MFCC5	MFCC8	MFCC10	Δ MFCC0	ΔΔ MFCC0	Loudness	ESPS Pitch
min	-0.77	-0.79	0.20	-0.78	-0.71	-0.59	-0.72	-0.79	-0.75	0.22	-0.02
max	0.23	0.16	0.82	0.36	0.47	0.59	0.54	0.78	0.75	0.78	0.54
range	**0.78**	**0.79**	**0.83**	**0.79**	**0.78**	**0.78**	**0.78**	**0.83**	**0.78**	**0.79**	**0.69**
mean	-0.56	-0.68	0.53	-0.48	-0.32	0.06	-0.11	-0.10	-0.07	0.57	0.30
std	0.66	0.71	0.67	0.69	0.62	0.63	0.68	0.66	0.63	0.69	0.55
skewness	-0.57	-0.57	0.07	-0.45	-0.45	-0.23	-0.39	0.40	-0.22	0.41	0.21
kurtosis	0.44	0.40	0.10	0.25	0.36	0.29	0.31	0.45	0.55	0.41	0.39
ZCR	-0.61	-0.67	-0.22	-0.52	-0.32	-0.43	-0.57	-0.22	-0.27	-0.10	-0.14
%25	0.59	0.62	-0.40	0.20	0.05	-0.02	0.00	-0.49	-0.41	-0.55	0.13

Strong correlation was also observed for maximal lower lip protrusion (D7) ($\rho = .60$). All these three measures received comparable strong correlations when computed as a mean for whole episodes. On the other hand these three distances correspond to the activation of the action units AU 25 and AU 26. This might suggest that the perceived degree of the intensity is correlated with the mean and maximal activation of AU 25/26 and, in other words, with the mouth opening. Similar relations were not observed for other action units that occur in laughter expressions. Indeed, in our test the correlation between the perceived intensity and the measures D4 and D5 was weak ($\rho = .33$ and .43). It suggests that the intensity of the orbicularis oculi activity (i.e. AU6) is not related to the perceived intensity. However it does not mean that this activity was not observed in the dataset. Similarly we did not observe a relation between the measurements corresponding to AU 12 and the perceived intensity. Indeed, the correlation between perceived intensity and the measurements D3, D8, and D9 was only slightly higher (0.33-0.48 for maximum functional, and 0.31 - 0.43 for mean functional) than for the distances corresponding to AU 6. Finally, frowning is even less correlated with the perceived intensity. The observed correlation for the maximal value of the measurement D12 is 0.37. The detailed data are presented in Table 2.

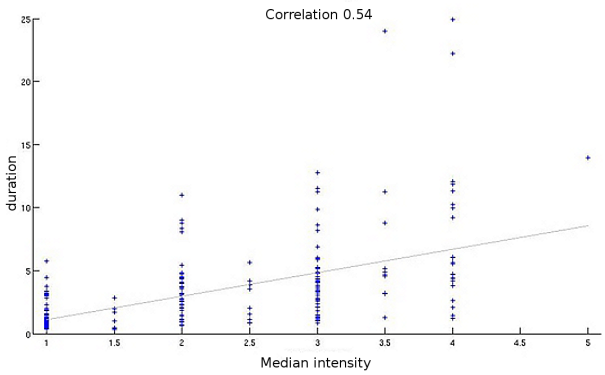

Figure 6: Correlation between median intensity and laughter duration

Interestingly, the overall duration of the laugh is not strongly correlated ($\rho = .54$) with the perceived intensity (Figure 6). In other words, an intense laugh does not necessarily last long, and vice-versa.

These results show us that some audio and/or visual features are strongly related to the perceived intensity of laughs. Hence these features are both good candidates to predict laughter intensity, and helpful to synthesize laughs with the desired intensity.

5.2. Intensity and respiration phases

In task 2 we studied the relation between the perceived intensity and the respiration phases. In total, the 249 laughs contain 419 exhalation phases and 190 inhalation phases. For each feature, we compare its distributions in inhalation and exhalation phases. A Lilliefors test showed that most of the features do not follow a Gaussian distribution; hence a Kolmogorov-Smirnov test was preferred to a t-test to compare the feature distributions over the 2 classes. The Kolmogorov-Smirnov test yielded in highly significant differences in the distributions of the 2 classes, for most of the audiovisual features. Figures 7 and 8 present the distributions, for the two classes, of 4 different features. These experiments illustrate that audiovisual features present different patterns in exhalation and inhalation laughter phases, which confirms our expectations since it is easy for humans to distinguish these phases. These features can be used for segmenting respiration phases in laughter and analyzing their differences.

6. Future works

In this paper we analyzed audio and visual features of spontaneous laughter expressive behavior. First of all we described the intensity annotation of an AVLC audiovisual corpus of spontaneous laughter. We also studied the relation between audio and visual cues of laughter and the perceived laughter intensity, as well as between the audio and visual features and laughter inhalation and exhalation phases.

Several limitations of this work should be noted. First of all the manual annotation of phase respirations can be only roughly done from the audio and/or visual channel. In future we plan to extend our work by using respiration sensor data to increase the segmentation accuracy. Secondly the referred results depend strongly on the choice of

Table 2: Correlation between laughter median intensity and the distances

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
min	0.39	0.32	0.12	0.19	0.11	0.17	0.26	0.15	0.19	-0.08	-0.10	0.08
max	**0.68**	**0.65**	0.48	0.28	0.30	0.20	**0.60**	0.43	0.33	0.12	0.01	0.37
range	0.52	0.48	0.46	0.43	0.44	0.40	0.54	0.3	0.36	0.19	0.15	0.23
mean	**0.64**	**0.61**	0.43	0.26	0.26	0.19	0.54	0.4	0.31	0.04	-0.03	0.29

the episodes, the segmentation method and the context in which the data were collected. Thus, we plan to use data from different video-corpora to confirm our results. It is particularly important to study the relation between the perceived intensity, some characteristics such as occurrence of AU6 and the type of laughter (social, hilarious). Last but not least the intensity annotation score corresponds to the whole episode but continuous annotation might be more informative as the intensity may not be constant during the laughter episode.

This is an ongoing work. Future works will consist in the more detailed annotation of the existing corpus, more detailed data analysis and finally building laughter models. First of all we plan to extend the intensity annotation of our video-corpus. We will annotate separately the audio and video channels using the same protocol as the one used in Section 3. We are particularly interested in the relation between the evaluation of the single modalities and the overall perception of the intensity. Taking into consideration that laughter episodes are often silent (at least in some phases), this work will give us more knowledge about the role of single modalities in laughter episodes.

Secondly, we are currently investigating the relation between facial actions and the produced laughter sounds, which will also help the synchronized audiovisual laughter synthesis, by looking at the relationship between the annotated vowel-like phones of the AVLC corpus and the shape of the mouth.

Thirdly, after finishing the annotation we discussed with some annotators about the task they had worked on. From these free discussions we observed that our annotators were often trying to evaluate laughter intensity in a subject-dependent way: they evaluated some laughs as relatively intense, i.e. intense when considering that specific person, even if they were not explicitly requested to do so. Our hypothesis is that, while coders may evaluate inter-subject intensity in the first episodes of laughter for a given subject, they rather evaluate the intra-subject intensity when the number of episodes increases. This hypothesis needs to be verified in future works. We ignore this factor in the analysis presented here.

Finally, the results presented here provide new insight for laughter synthesis. We have a better idea of how audiovisual features are related to laughter intensity and respiration phases. We can also use these results for actual prediction of laughter intensity and segmentation of inhalation and exhalation phases.

7. Acknowledgements

The authors would like to thank all volunteer annotators of the database. This work was supported by the European FP7-ICT-FET project ILHAIRE (grant n270780).

8. References

V. Adelsward. 1989. Laughter and dialogue: The social significance of laughter in institutional discourse. *Nordic Journal of Linguistics*, 102(12):107–136.

P. Drack, T. Huber, and W. Ruch. 2009. The apex of happy laughter: A facs-study with actors. In E. Banninger-Huber and D. Peham, editors, *Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions*, pages 32–37. Word Scientific Publisher.

T. Drugman, J. Urbain, and T. Dutoit. 2011. Assessment of audio features for automatic cough detection. In *19th European Signal Processing Conference (Eusipco11)*, pages 1289–1293, Barcelona, Spain, August 29 - September 2.

P. Ekman and W. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto.

T. Huber, P. Drack, and W. Ruch. 2009. Sulky and angry laughter: The search for distinct facial displays. In E. Banninger-Huber and D. Peham, editors, *Current and Future Perspectives in Facial Expression Research: Topics and Methodical Questions*, pages 38–44. Word Scientific Publisher.

K. Krippendorff. 2012. Computing Krippendorff's alpha-reliability. http://www.asc.upenn.edu/usr/krippendorff/dogs.html (last accessed February 16 2012).

G. Peeters. 2003. A large set of audio features for sound description (similarity and classification) in the cuidado project. Technical report.

W. Ruch and P. Ekman. 2001. The expressive pattern of laughter. In A.W. Kaszniak, editor, *Emotion qualia, and consciousness*, pages 426–443. Word Scientific Publisher.

K. Sjölander and J. Beskow. 2011. Wavesurfer: open source tool for sound visualization and manipulation [computer program]. http://sourceforge.net/projects/wavesurfer/.

J. Urbain and T. Dutoit. 2011. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In *Proceedings of the fourth international conference on Affective Computing & Intelligent Interaction*, pages 397–406, Memphis, Tennessee, USA. Springer-Verlag.

J. Urbain, E. Bevacqua, T. Dutoit, A. Moinet, R. Niewiadomski, C. Pelachaud, B. Picart, J. Tilmanne, and J. Wagner. 2010. The AVLaughterCycle Database. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*,

pages 2996–3001, Valletta, Malta. European Language Resources Association (ELRA).

J. Wagner, E. André, and F. Jung. 2009. Smart sensor integration: A framework for multimodal emotion recognition in realtime. In *Proceedings of the third international conference on Affective Computing & Intelligent Interaction*, pages 1–8, Amsterdam, Holland.

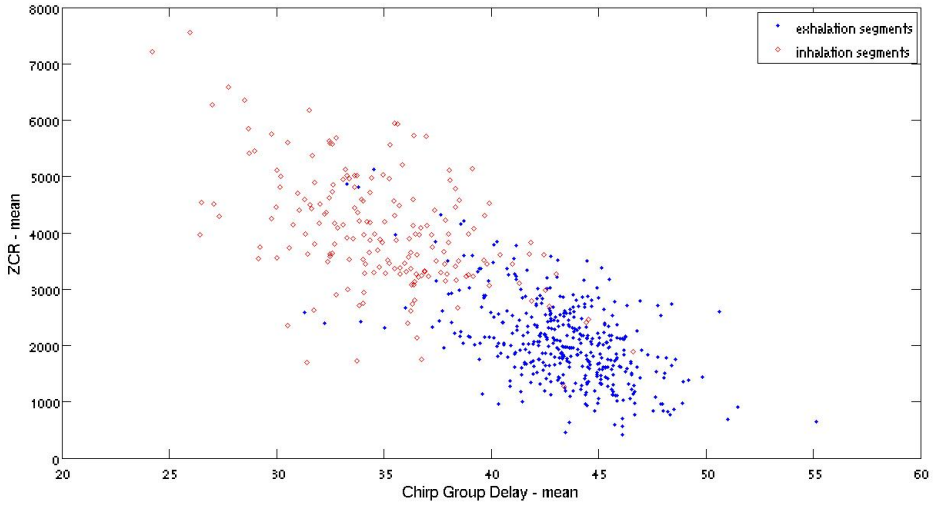

Figure 7: Distribution of mean Chirp Group Delay and mean Zero-Crossing Rate for exhalation and inhaltion laughter phases

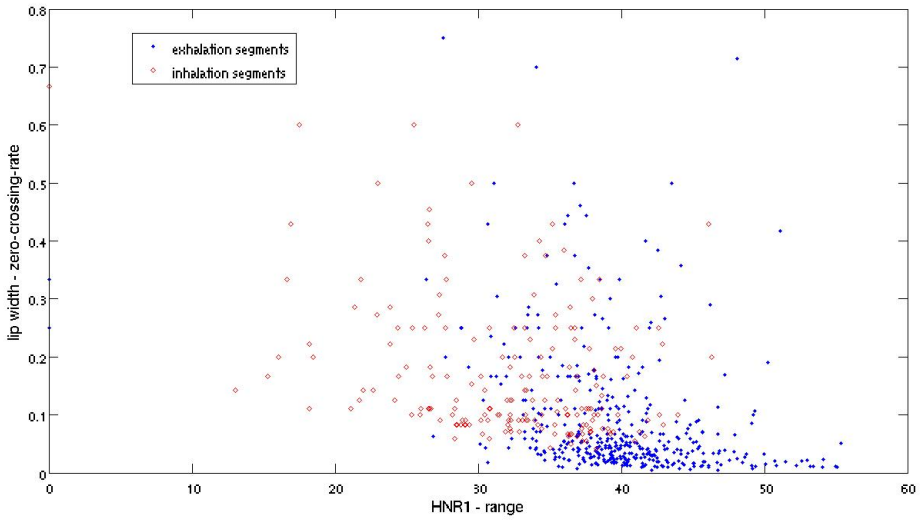

Figure 8: Distribution of mean HNR1 range and zero-crossing rate of AU6 for exhalation and inhalation laughter phases

ILHAIRE Laughter Database

Gary McKeown[1], Roddy Cowie[1], Will Curran[1], Willibald Ruch[2], Ellen Douglas-Cowie[1]
[1]School of Psychology, Queen's University, Belfast, UK
[2]Department of Psychology, University of Zürich, Switzerland
E-mail: g.mckeown@qub.ac.uk

ILHAIRE Laughter Database—Phase 1

The ILHAIRE project seeks to scientifically analyse laughter in sufficient detail to allow the modelling of human laughter and subsequent generation and synthesis of laughter in avatars suitable for human machine interaction. As part of the process an incremental database is required providing different types of data to aid in modelling and synthesis. Here we present an initial part of that database in which laughs were extracted from a number of pre-existing databases. Emphasis has been placed on extraction of laughs that are social and conversational in style as there are already existing databases that include instances of hilarious laughter. However, an attempt has been made to exhaustively extract all instances of laughter from databases that were not designed for the purpose of generating hilarious laughter. Theses databases are: the Belfast Naturalistic Database, the HUMAINE Database, the Green Persuasive Database, the Belfast Induced Natural Emotion Database and the SEMAINE Database.

Keywords: Laughter, Database, Social

1. Introduction

The ILHAIRE project seeks to scientifically analyse laughter in sufficient detail to allow the modelling of human laughter and subsequent generation and synthesis of laughter in avatars suitable for human machine interaction. As part of the process an incremental database is required providing different types of data to aid in modelling and synthesis. The database is termed "incremental" as different teams within the project require different types of data and at varying stages during the life of the project. At the end of the project there will be a substantial database, which will contain laughs extracted and annotated from existing databases in addition to the generation of specific laughter material. The latter will be recorded and annotated in detail, using FACS annotation and motion capture data of both facial features and full body motion during laughter events.

The project recognises that laughter includes not only hilarious laughter but also various forms of social laughter. Hilarious laughter occurs typically in reaction to a stimulus such as a joke or a funny video. Importantly, it can occur either when the laugher is alone, or in the presence of others. Social laughter, however, only occurs during social interactions typically in conversations involving two or more participants. It is thought to serve several functions in conversations: it can regulate a conversational interaction, alter the meaning of an utterance, provide a backchannel signal that acknowledges engagement in the conversation, or signal a level of group cohesion (Vettin & Todt, 2004).

There are already existing databases dedicated to providing instances of laughter: including the AVLaughterCycle database produced by members of the ILHAIRE project, which will not be reported on here; as well as the MAHNOB laughter database (Petridis et al., In Press). These databases focus primarily on hilarious laughter. To provide a preliminary overview of laughs that are more social in nature, we have extracted laughter from five existing databases designed to show people acting and interaction in a variety of situations that are relatively natural, but emotionally coloured. Five databases were chosen to extract these kinds of laugh. Laughter was not a criterion in the construction of any of these databases, and so there is no bias either towards the presence of laughter, or towards the presence of any particular type of laughter, in them. As this is an initial attempt to extract and social and conversational laughter multiple naturalistic databases were used with the goal of an exhaustive search extracting laughter where it was observed to occur and often in natural settings not typically associated with laughter. The database reported here contains all the extracts from these databases which did contain laughter, with associated labels. An attempt was made by one person to exhaustively extract laughter from these databases. This was followed by further validation of a subset of the extracted laughs. This paper reports only on the initially extracted laughs the validation will be reported in greater detail a future paper. It will be made available as part of the broader ILHAIRE database. The nature of data collection in each of these databases is explained in greater detail in the references associated with the original databases.

The paper will introduce each of the databases that were used in the creation of this initial phase of the ILHAIRE Laughter Database and address the issues that arise due to the idiosyncrasies of the original database. This will be followed by details of annotations that are available and future annotation plans.

2. Belfast Naturalistic Database

The Belfast Naturalistic Database (Douglas-Cowie, Campbell, Cowie, & Roach, 2003) was an early attempt to gather a broad swathe of audio-visual material of people who at least appeared to be experiencing genuine

emotion. These were primarily drawn from television programmes, talk shows, religious and factual programmes. The material contains a broad sample of both negative and positive emotions, with 53 of the total of 127 video clips containing laughter in some form. There are copyright issues associated with many of the video clips in the Belfast Naturalistic Database which unfortunately means that only five of the clips can be broadly disseminated with the ILHAIRE Laughter Database.

3. HUMAINE Database

The HUMAINE database (Douglas-Cowie et al., 2007) was created with the purpose of demonstrating the breadth of material that exists related to a broad understanding of the word emotion—termed 'pervasive emotion'. The database contains fifty audio-visual clips from a variety of sources providing diverse examples of emotional content relevant to affective computing. From these fifty clips 46 instances of laughter were extracted for inclusion in the current database. The quality of these clips is variable, but they are useful as illustrations of the variety of situations in which laughter occurs.

4. Green Persuasive Database

The Green Persuasive Database (Douglas-Cowie et al., 2007) contains a collection of audiovisual clips that were recorded to capture a type of interaction where there are strong feelings, but not basic emotions. The scenario involves one participant who is trying to convince the other participant of the moral case for trying to adopt a more environmentally friendly lifestyle, using as examples sustainable transport, flying less, and reducing greenhouse gas emissions. The conversations are mildly confrontational but persuasive and friendly rather than overtly argumentative. There is a strong power imbalance between participants as the persuader is a University Professor and the listeners are students. There were eight interactions in total lasting between 15 and 35 minutes. From these eight participants 280 instances of laughter were extracted. The nature of the interactions meant that most of these laughs are conversational or social laughs that occur as a natural part of a social interaction between two people. Very few would be classified as hilarious laughs.

5. Belfast Induced Natural Emotion Database

The Belfast Induced Natural Emotion Database (BINED) (Sneddon, McRorie, McKeown, & Hanratty, 2012) represents a deliberate effort to induce specific kinds of emotional behavior. The goal of the database was to produce material that could act as replacements to the posed static photographs that are often used in studies of emotion. Natural dynamic emotion was elicited either by watching emotional video clips or by a series of tasks in which participants actively engaged. The database is organized into three sets based around chronological data collection periods. The first set involved tasks designed to elicit: Amusement, Disgust, Fear, Frustration, and Surprise. There are 113 participants, 43 females and 70 males in Set 1 of the database. Laughs have been extracted from this set, 289 instances of laughter were extracted from a total of 565 clips. These occurred at different frequencies depending on the kind of emotion that the task sought to elicit. Figure 1 plots the frequency of the laugh instances for males and females in these clips for each of the different tasks. Importantly the number of clips differs and so this information serves only to display the numbers of instances of each clip in the database and not a comparison of levels of laughter in each gender. Work is ongoing to add laughs from Set 2 and Set 3 of the database. This work includes the extraction of laughs from the Amusement clips by 9 raters, and will provide a greater reliability to the laugh extraction as well as providing some knowledge about the ambiguity involved in deciding where the exact onset and offset points are in a given laugh. While laughter onset is typically the easier of these to distinguish, identifying onset can be particularly challenging when laughter is preceded by a smile; knowing how and when a smile becomes a laugh is an open question. Greater challenges are posed in identifying laughter offset, this can often be further compounded by a second bout of laughter can occurring before there is a return to a neutral face. We hope to address some of these issues with the clips from BINED.

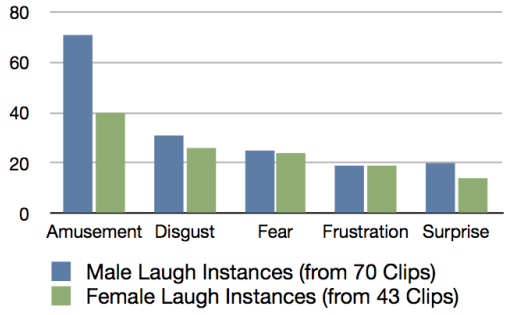

Figure 1: Laughter Instances for the task types in the BINED database

6. SEMAINE Database

The SEMAINE database (McKeown, Valstar, Cowie, Pantic, & Schröder, 2012) provides high quality audio-visual clips from a setting that is strongly aligned with the goals of the ILHAIRE project. The SEMAINE project developed a system which engages users in a sustained emotionally coloured interaction with an avatar—known as a Sensitive Artificial Listener (SAL)—with a primary emphasis on creating technology that attended to and synthesized the non-verbal components of human interaction. As part of the project there were different stages of interaction in which the various levels of engagement approached the end goal of a human machine interaction. In the initial stage known as Solid SAL one participant took the role of the user

and another took the role of the avatar and acted as one of the four SAL characters in the SAL system. A later version—Semi automatic SAL—used a human interacting with an avatar operated by another human; users could see only a schematic image of a face and the operator selected pre recorded utterances from a set script. The final stage involved interactions with an autonomous avatar controlled by the fully automatic SAL system. Once again there was no explicit remit within the SEMAINE project that called for laughter in the interactions, the laughter that occurred was largely conversational and social laughter incidental to the task of interacting with the avatar or with a person pretending to be an avatar. The laughs that are included in this version of the ILHAIRE database are taken from the Solid SAL interactions and therefore involve interaction between two humans. Laughs were automatically annotated by the audio feature recognition components of the openSMILE system within the SAL system and then extracted using this annotation. While these laughs were checked by a human and some false positives were removed, it is possible that some laughs that were not recognized by the system and, therefore, the list cannot be considered an exhaustive extraction of the laughs in the Solid SAL section of the SEMAINE database. In total 443 instances of laughter were extracted from 345 video clips.

Expressing predominantly positive emotions		**Mean Rating Frequency**
Happy	Indicative of pleasure, contentment or joy because of a particular thing	**16.3**
Relieved	Laughter that signifies a concern or anxiety has been laid to rest	**5.9**
Thankful	Appreciative or expressing gratitude to a person	**5.1**
Hilarious	Unrestrained response to something that is simply found extremely funny	**7.2**
Giggling	Happy mixed with a sense that something is ridiculous or unimportant	**13.8**
Helpless	Laughter that is positive, but that the person has completely lost control of	4.9
Lustful	Salacious: expresses sexual arousal and anticipation	0.1
Mischievous	Playful but with the intent to case trouble	2.4
Boastful	Laughing out of self-admiration or excessive pride	3.1
Expressing predominantly negative emotions		
Angry	Laughter that conveys aggression or intimidation	0.3
Sarcastic	Laughter to convey that words spoken should be taken as cynical or mocking	1.4
Contemptuous	A laugh that expresses superiority to the person being laughed at, showing disdain or scorn towards them	1.0
Sullen	Laughter that indicates someone is being pressurised to behave in ways that he/she resents	1.4
Tense	Uncomfortable laughter. Used in situations where it is unsure what should be said	**6.4**
Embarrassed	A result of being self-conscious or an expression of confusion or shame	**9.1**
Hysterical	Laughter that the person has completely lost control of as a result of feeling that he/she has lost control of events.	1.7
Desperate	Frenzied laughter conveying a dire need for something	1.2
Sad	Laughter indicating regret that something has happened, with resignation that it cannot be changed	2.8
Expressing emotions with positive and negative elements		
Shy	Nervous and quiet, trying not to make feelings too obvious	4.2
Anxious	Experiencing unease and trying to lessen it with laughter	**10.9**
Apologetic	Awkward laugh used when an individual is trying to express an apology / show remorse.	1.5
Meaningful	A laugh to show there is more meaning to what has been said than is simply expressed	3.6
Cunning	Laughter that is shrewd, sly or deceitful	0.4
Taunting	Laughter directed at someone in particular. Intended to make fun of or belittle him/her.	1.8
Schadenfreude	Laughter expressing pleasure in the misfortune of another person	1.8
Other - not primarily expressing emotions		
Physical reflex	Response to physical prompt (usually tickling)	0.4
Surprised	A reaction to astonishment, when something happens suddenly or unexpectedly	**8.4**
Backchannelling laughter	Laughter that is part of a conversation, and conveys a routine acknowledgement of what the other speaker has just said	**8.3**
Polite laughter	Laughter aimed at being courteous and showing good manners	**7.9**
Contrived	A forced or planned laugh	**5.2**
Staged laughter	Completely forced. Usually detectable easily. The kind of laugh found in TV/films	0.9
Other	Not included in the list above	2.0

Table 1. Classification Scheme for Annotation of Laughter

7. Annotation

There are existing annotations that are provided with the databases, which were collected for their original uses. These give mainly information about emotional content. To these ongoing projects are adding annotations of the laughter in the databases. Here we will comment only on the laughter-specific annotations. The attempt to annotate the onset and offset of amused clips in the Belfast Induced Natural Emotion Database has already been outlined. Additional to the goal of establishing inter-rater agreement of onset and offset times this data can be used to establish duration of laughter, and raters have also been asked to produce a rating of the intensity of laughter on a scale between 1 and 10.

A second associated project has attempted to classify the types of laughter using the clips found in the Belfast Naturalistic Database and HUMAINE database. Starting with an initial classification of 23 laughter types (Drack & Ruch, 2007) this was extended to the laughter classification scheme that can be seen in Table 1. The descriptions in each category were developed in conjunction with users to ensure that they could be readily understood by non-experts. 16 raters have classified the clips using these categories. The final column in the table shows the average number of times each label was used per rater, and so it gives a broad indication of the frequency with which different kinds of laughter appear in a pre-existing body of naturalistic material. This is not conclusive, but it gives a first indication of the kinds of laughter that should be a priority for research concerned with facilitating interaction.

The broader annotation strategy of the database is a yet undetermined. Where available resources are used for annotation will be decided depending on the outcome of the preliminary annotation research such as that outlined in Table 1, and the general requirements of the members of the ILHAIRE project.

8. Future Development

The database detailed in this paper has been developed as an initial phase in an incremental database which is being created as part of the IHAIRE project. These initial components will be added to with laughter data specifically collected according to the needs of the project. This will include full body motion capture data with accompanying audiovisual data and face only motion capture with accompanying audiovisual data. The goal is to collect a broad variety of types of laughter within the broad categories of social and hilarious and more refined categories outlined in Table 1. As this data is collected and annotated it will become part of the ILHAIRE database and be made available to the research community.

9. Availability

We plan to make the database available for use by the broader research community in the near future.

10. References

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

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., McRorie, M., Martin, J., Devillers, L., Abrilian, S., Batliner, A., Amir, N. & Karpouzis, K. (2007). The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. Lecture Notes In Computer Science, 4738, 488–500.

Drack, P., & Ruch, T. H. W. (2007). The Apex of Happy Laughter: A FACS-Study with Actors. In E. Banninger-Huber & D. Peham (Eds.), Current and Future Perspectives in Facial Expression Research: Topics and Methodological Questions 32–37. Innsbruck University Press.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schröder, M. (2012). The SEMAINE Database: Annotated Multimodal Records of Emotionally Coloured Conversations between a Person and a Limited Agent. IEEE Transactions on Affective Computing, 3(1). doi:10.1109/T-AFFC.2011.20

Petridis, S., Martinez B., & Pantic, M. (In Press) "The MAHNOB-Laughter Database", Image and Vision Computing Journal.

Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The Belfast Induced Natural Emotion Database. IEEE Transactions on Affective Computing, 3(1). doi:10.1109/T-AFFC.2011.26

Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. Journal of Nonverbal Behavior, 28(2), 93–115. Springer.

Comparing Non-Verbal Vocalisations in Conversational Speech Corpora

Jürgen Trouvain[1] & Khiet P. Truong[2]

[1]Saarland University, Germany & [2]University of Twente, The Netherlands

E-mail: [1]trouvain [at] coli.uni-saarland.de & [2]k.p.truong [at] utwente.nl

Abstract

Conversations do not only consist of spoken words but they also consist of non-verbal vocalisations. Since there is no standard to define and to classify (possible) non-speech sounds the annotations for these vocalisations differ very much for various corpora of conversational speech. There seems to be agreement in the six inspected corpora that hesitation sounds and feedback vocalisations are considered as words (without a standard orthography). The most frequent non-verbal vocalisation are laughter on the one hand and, if considered a vocal sound, breathing noises on the other.

1. Introduction

Conversations do not only consist of spoken words but they also consist of non-verbal signals transmitted via the acoustic channel. Typical of these signals are that they often do not appear in dictionaries which is one of the reasons why people often have trouble writing down the signal's sound in orthographical form. Examples of these signals are laughter, coughs, breath sounds and feedback sounds such as "hmm-mm". We call these signals Non-Verbal Vocalisations (NVVs). Some of these vocalizations clearly have a communicative function and some are a result of the planning processes of speech production (what am I going to say next and how am I going to say it). As a consequence, NVVs are generally more present in spontaneous (conversational) speech than in carefully read aloud speech.

Research on NVVs in spontaneous conversational speech has been limited, which is partly due to the fact that NVVs are usually considered non-speech or 'garbage' sounds, especially from a technology point of view. Traditional automatic speech recognition (ASR) systems usually discard NVVs as non-speech sounds. However, researchers are becoming more aware of the importance of NVVs in spontaneous conversational speech and the need to model NVVs. Nowadays, ASR systems need to be able to recognize conversational speech and cope with NVVs. In addition, it is known that NVVs can carry communicative and affective meaning, which can be modelled for the development of spoken dialogue systems and emotion-aware systems.

Another possible reason for the limiting research performed on NVVs concerns the huge variability of NVVs. There is no clear definition of NVVs and there are no standard transcription and annotation protocols. These issues may have discouraged researchers to investigate NVVs in depth. Previous work on NVVs includes Ward (2006) in which a description of so-called conversational grunts in American English is presented. The focus of that study seems to cover only a part of the NVVs by our definition. We take on a broader view and include vocalizations such as laughter and audible breath sounds, which could play a role in dialogue. Our aim in this paper is to shed some light on the variability of NVVs.

The descriptive aims of study are to present various types of NVV and to sketch a scheme to structure various NVVs. The analytical part is to check which categories of NVV were considered in the different corpora and to find out i) differences in usage of NVV labels between different corpora and ii) frequencies of occurrence of various annotated NVV types. The results allow us to identify why and which NVVs can be important for communication research in conversational speech, and hence should be annotated with higher priority.

2. Types of non-verbal vocalisations

One problem of grouping and classifying NVVs is that the same or similar phonetic token can represent different NVVs. Breath intakes for example can be observed either as a vegetative sound or as part of a laugh or as a pragmatic signal with the meaning "I would like to have the turn." Here, we describe a number of possible types of NVVs.

2.1 Vegetative sounds

Vegetative sounds are not primarily communicative and not all are under voluntary control. Examples include snoring, moaning (e.g. in sports), swallowing sounds, chewing noises (with open or closed mouth), hiccup, coughing, sneezing, clearing the throat, yawning or panting (after physical exercise). Typically, vegetative sounds are not learned. However, there are vegetative sounds that require some level of learning such as spitting (e.g. cherry stones), lip smacking or producing an ingressive [s]. Probably the most frequent vegetative sound is audible inhalation. Audible exhalation sounds will also occur in conversation (not only after physical exercise).

Vegetative sounds can be used deliberately like clearing the throat ("ehem") to "say" that e.g. "I'm here now". Thus, deliberate vegetative sounds require pragmatic knowledge and the control of the vocal apparatus.

2.2 Affect sounds

Affect sounds include vocalisations such as laughing, weeping, cheering, crying loud and screaming. Conventionalised forms of these affect sounds include the deliberate use of moaning and yawning as well as imitations of coughing and snoring.

Schröder (2003) uses the more known term affect burst for affect sound, but then in a broader sense. It goes beyond the just described affect sounds incorporating also interjective words like "yippie" and "igitt".

2.3 Interjections as 'semi-words'

Sometimes the term interjection is used to indicate all kinds of NVVs with a paralinguistic character or at least those which "are tied to emotional or mental attitudes or states" (Wharton 2003). Sometimes interjections are meant to represent a certain word class, which would make them verbal vocalisations. Their debated linguistic status, the frequently unclear orthography and the fact that they often are not listed in dictionaries make them candidates for 'semi-words' (Wharton 2003).

Although there is no generally accepted definition of interjections they are often divided into primary and secondary interjections. The latter are words with an own meaning like "Damned! " or "Shit!" making them clearly verbal vocalisations. Primary interjections are e.g. "ouch" or "wow".

Onomatopoetic expressions like "miaow", "cuckoo", "knock-knock" can also be analysed as primary interjections, however, without any affective component. This is in contrast to interjections imitating environmental sounds in a less conventionalised way such as "woosh" or "bing". A further sub-category of primary interjections are affective words with an ungrammatical phonology such as "pst" or "shh" (no vowels) and "ts-ts-ts-ts" (clicks).

2.4 Feedback and filler sounds as 'semi-words'

Other 'semi-words' but without any affective component are hesitation sounds, also known as fillers or filled pauses such as "uh" or "uhm". Often they are regarded as disfluencies to which lengthened syllables (or syllable drawls) can be counted as well although this lengthening effect is not an independent vocalisation.

Another category of "semi-words" are sounds which function as feedback signals. They include humming signs like "hm" or "yeah" and "uhu". Usually they are used to backchannel but potentially also for asserting and other kinds of attitudinal expression.

2.5 Melodic utterances

A universal phonetic behaviour is the use of melodies with the own vocal apparatus. Melodies without text can be hummed, sung or whistled. We do not expect many of these utterances in conversation.

3. Distinctive dimensions

The same phonetic expression can be used for various functions. For instance breath sounds are primarily vegetative sounds. But breathing noises also play a role for laughter. Also an affect sound signalling startle usually involves a strong and sudden inhalation. Furthermore, audible inhalation can be used to signal to take the turn in a conversation. Another example is the humming sound (or neutral nasal consonant) which can be used for melodic purposes as well as for feedback signals and also for affective sounds signalling disgust but also pleasure – depending on its voice quality and its prosody. For this multi-functionality of NVVs we propose to describe them along four various distinctive dimensions of which one is binary ('vegetative') and three are not meant to be binary but continuous.

3.1 Vegetative dimension

Not all NVVs have a paralinguistic character and are uttered by the speaker to transport information. However, they contain extra-linguistic information about the speaker that can normally not be changed, e.g. coughs and sneezing can signal the status of the health or coughs can also be used for recognising the identity of a speaker. The vegetative dimension includes also not explicitly vegetative NVVs without any communication partner, e.g. affect sounds expressing pain.

3.2 Spelling dimension

There is no clear-cut border between NVVs in a narrow sense and semi-words. The decisive dimension to consider a vocalization as belonging to one of the semi-word classes or not seems to be the spelling dimension. Several times a continuum has been proposed reaching from 'raw' affect bursts (cf. Schröder 2003) or 'natural sounds' (cf. Wharton 2003) at the one end and secondary interjections at the other. At the one extreme reliable spelling of the expressed sounds is (nearly) impossible, on the other extreme the orthographic standard is rather clear. The spelling dimension also reflects the fact that NVVs at the non-spelling end are phonetically encoded by glottal rather than supra--glottal activities.

3.3 Affective dimension

Affect sounds and (most) interjections are defined by the affective dimension thus transporting a lot of information about the speaker and her/his attitudes and feelings in a very short time. Affective information is usually not present with vegetative sounds and filler sounds. Feedback sounds, however, can sometimes transport affective information.

3.4 Pragmatic dimension

Some NVVs act as pragmatic particles with functions for the management of the conversation. For instance feedback sounds such as 'backchannels' are indispensable for keeping a conversation fluent. Filler sounds can signal some problems with the self-management of the talker but it can also show upcoming new information. But also laughter and other affect sounds can be used as a feedback signal.

A summarization of the types of NVV as described in section 2 and the proposed dimensions in this section can be found in Table 1. It must be noted that the classification presented in types and dimensions is just a *sketch* for further theoretical considerations as well as empirical analyses.

Table 1: Gray areas and plus-signs indicate the presence and the intensity of the three continuous dimensions for the various types of NVVs (the binary dimension 'vegetative').

dimensions / types	veg.	spelling	affective	pragm.
vegetative sounds		-	-	-/+
deliberate veget. s.		+	+	+
affect sounds		-	+++	-
deliberate affect s.		+	++	-
imitative sounds		+	+++	-
melodic utter.		-	+++	-
interjections		++	++	-
fillers		++	-	+++
feedback sounds		++	-/+	+++

4. Differences in usage

Six different corpora of conversational English were inspected: ICSI meeting corpus (Janin et al. 2003), AMI (Carletta 2007), Switchboard (Godfrey & Holliman 1997), Diapix Lucid corpus (Baker & Hazan 2011), HCRC Map Task corpus (Anderson et al. 1991) and the Buckeye corpus (Pitt et al. 2007).

Annotations of the above mentioned NVV widely differ among corpora of conversational speech. All corpora consider the "semi-words" listed in sub-section 2.3 as words, although the orthography differs very much. It must be noted that a comparison is very hard due to different treatments of NVV annotations as tokens in the various annotation schemes but also due to various annotators, differences in conversational tasks and differences in microphones.

Laughter is always annotated in the corpora under inspection. However, speech-laughs were not always annotated as such (see table 2). Despite the various differences of the inspected corpora it seems obvious that annotated "laughs" is the predominant type of NVV in all corpora (cp. Fig. 1): more than 60% of all annotated NVVs in AMI and more than 40% in ICSI and Switchboard. However, the remaining three corpora show a remarkably low number of laughs, which can be attributed to a smaller amount of recorded data, the dyadic of multiparty character, and the conversational task.

The differences regarding breathing sounds are rather dramatic (see fig. 1). In the Buckeye corpus breath sounds are not a category at all whereas in AMI the transcription guidelines provide an appropriate annotation tag but it was extremely rarely selected (0.2% of all NVVs).

We understand that breath sounds in Switchboard were treated as the 'other' category which was named 'noise'.

Table 2: Table of occurrences of NVVs in various corpora. ‘N/A’ means that the vocalization was not explicitly mentioned in the transcription guidelines and was hence not considered by the transcribers. A zero ‘0’ means that the vocalization was mentioned in the transcription guidelines (and thus considered by the transcribers) but we cannot count these because there were not any or they were included in an explicit ‘Other’ category. ‘The rest’ means all the other annotated NVVs that did not fit one of our categories under inspection.

	Multiparty				Dyad							
	ICSI		AMI		Switchboard		Diapix		HCRC		Buckeye[1]	
N conversations	75		171		2438		57		128		255	
Duration	72h		100h		518h		7.3h		14.5h		37.8h	
	Abs	%	Abs	%	Abs	%	Abs	%	Abs	%	Abs	%
Laugh	12643	40.8	16477	61.0	22209	37.4	582	8.9	1002	5.3	1899	7.2
Speech-laugh	1017[2]	3.3	n/a	n/a	13503	22.7	333	5.1	n/a	n/a	1020	3.9
Breath	12465	40.2	57	0.2	0	0	3539	54.2	12280	64.8	n/a	n/a
Cough	256	0.8	1114	4.1	0	0	n/a	n/a	320	1.7	0	0
Clearing the throat	906	2.9	0	0	0	0	0	0	n/a	n/a	0	0
Lip smacking	n/a	n/a	3	0.0	n/a	n/a	1182	18.1	4512	23.8	n/a	n/a
Eating	39	0.1	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Yawn	62	0.2	10	0	0	0	n/a	n/a	n/a	n/a	n/a	n/a
Sigh	22	0.1	47	0.2	0	0	0	0	n/a	n/a	0	0
Humming/Singing/ Whistling	47	0.2	85	0	0	0	n/a	n/a	n/a	n/a	n/a	n/a
Other	n/a	n/a	8888	33.0	23682	39.9	893	13.7	n/a	n/a	22661	86.3
The rest	3554	11.5	310	1.1	0	0	0	0	823	4.3	685	2.6
Total	31011	100	26991	100	59394	100	6529	100	18937	100	26265	100

[1] Only one person of the dyad was recorded and annotated

[2] Counts of segments (instead of separate words) spoken while laughing

Diapix and HCRC show the expected high number of breath sounds whereas ICSI shows a medium-scaled number. This rather disparate picture is also reflected in the plethora of the often detailed tags such as "inbreath", "outbreath", "long loud outbreath", "loud inhale", "strong exhale" etc.

When looking at the token frequency of selected NVV types it can be easily observed that laughter and breathing sounds dominate. Other NVVs like cough, clearing the throat, yawning etc (see table 2) show a rather low frequency of occurrence (with the exception of lip smacking for the HCRC map task corpus).

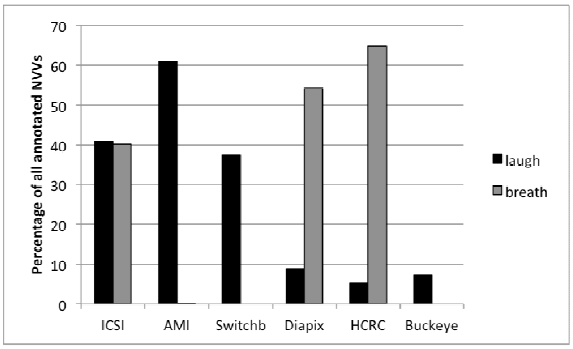

Fig1: Laughter and breathing sounds were the two main NVVs annotated in the inspected corpora. The graph shows the numbers of annotated "laugh" and "breath" relative to the total number of NVV for each corpus.

5. Concluding remarks

Our analysis of six corpora with conversational speech revealed that there is a *huge* disparity among the inspected corpora with regard to the annotation of NVVs. There seems to be agreement that 'semi-words' like feedback and filler sounds as well as interjections with a possible spelling should not be regarded as 'non-verbal' or 'non-speech'. It also turned out that laughter represents the category of NVV with the highest frequency of occurrence. However, there is disagreement about the status, the amount and the specification of breathing sounds. Other types of NVV such as coughing, eating sounds, yawning or melodic utterances either play only a minor role or are not yet explored in the inspected corpora of conversational speech. Although one could say that these sounds do not seem to be much dialogue-related, we do not recommend exclusion of these, as some of these sounds *can* be useful for dialog research. For example, a cough can contain speaker identity information and yawning or singing can be signals of tiredness or good mood.

Usually the details of the annotation of NVV depend on the goal of investigator's research. However, corpora of conversational speech provided for general research on how spoken interaction unfolds would also need a more detailed annotation of NVVs. Based on our investigations and with respect to future research we consider it worthwhile to have more consistent and detailed NVV annotations. In particular, research on turn-taking could benefit from consistent annotation of breath sounds which can also serve as additional signals for prosodic breaks in general.

The difficulty providing practically useful and theoretically valid definitions of NVV reflects the lack of knowledge about the acoustics as well as about the functions NVVs can serve. Some NVVs show similar phonetic shapes but serve different functions. For example, a schwa-sound or a neutral nasal consonant can occur as a token of each NVV type. It just depends on the glottal and sub-glottal activity (voicing, voice quality, intonation, respiration) *and* the context (syntactic position and articulatory isolation) that makes this sound have a certain interpretation. An analysis of additional annotations such as dialogue act annotations in which pragmatic functions like feedback, filler etc. are annotated could be helpful.

In order to provide a better basis for comparing different corpora a re-annotation of the NVV would be advisable. This would require a theoretical framework to put NVVs into a larger context of which here only a few points were discussed. A theoretical fundament backed with empirical data would also allow comparisons of NVVs between taken from experimental lab studies and spontaneous conversations.

Acknowledgements

This research has been supported by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet) and the UT Aspasia Fund.

6. References

Anderson, A.H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weintert, R. (1991). The HCRC Map Task Corpus. *Language and Speech* 34(4), pp. 351-366.

Baker, R, Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods* 43(3), pp. 761-770.

Carletta, J.C. (2007). Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation* 41(2), pp. 181-190.

Godfrey, J.J., Holliman, E. (1997). *Switchboard-1 Release* . Linguistic Data Consortium, Philadelphia.

Janin, A., Baron, D., Edwards, D., Ellis, D., Gelbart, D., Morgan, N. (2003). The ICSI meeting corpus. *Proceedings of ICASSP*, pp. 364-367.

Pitt, M.A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release) [www.buckeye corpus.osu.edu] Columbus, OH: Ohio State University. (Distributor).

Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication* 40 (1-2), pp. 99-116.

Ward, N. (2006). Non-lexical conversational sounds in American English. *Pragmatics and Cognition*, 14(1), pp. 129-182.

Wharton, T. (2003). Interjections, language and the 'showing'-'saying' continuum. *Pragmatics and Cognition*, 11(1), pp. 39-91.

Smiling Virtual Characters Corpora

Magalie Ochs[1], Paul Brunet[2], Gary McKeown[2], Catherine Pelachaud[1] and Roddy Cowie[2]

[1]CNRS-LTCI, TélécomParisTech, France
[2]Queens University, Belfast, UK
{ochs;pelachaud}@telecom-paristech.fr
{p.brunet;g.mckeown;r.cowie}@qub.ac.uk

Abstract

To create *smiling* virtual characters, the different morphological and dynamic characteristics of the virtual characters smiles and the impact of the virtual characters smiling behavior on the users need to be identified. For this purpose, we have collected two corpora: one directly created by users and the other resulting from the interaction between virtual characters and users. We present in details these two corpora in the article.

Keywords: Smiles, virtual agent, social stances

1. Introduction

To create *smiling* virtual characters, several issues need to be addressed. First of all, the different morphological and dynamic characteristics of the virtual characters smiles need to be identified. Indeed, a smile may convey different meanings such as amusement and politeness depending on subtle differences in the characteristics of the smile itself and of other elements of the face that are displayed with the smile (Ambadar et al., 2009; Ekman et al., 1982; Keltner et al., 1995). Moreover, a smile displayed by a virtual character may impact the users perception differently both positively and negatively - depending on when the virtual character expresses which types of smile (Krumhuber et al., 2008 ; Theonas et al., 2008). The second issue is then to identify the impact of the virtual characters smiling behavior on the users. To respond to these two issues, we have collected two corpora. The first corpus aims at collecting information on the virtual characters morphological and dynamic characteristics. The second corpus has been collected to measure the effects of virtual characters smiling behavior on users. We describe each of these corpora in the following.

2. Users-created corpus of virtual characters smiles

2.1. Tool to collect smiles

In order to identify the morphological and dynamic characteristics of the amused and the polite smile of a virtual character, we have proposed a *human-centric approach*: we have collected a corpus of context-free virtual characters smiles directly created by the users. For this purpose, we have created a web application (called *E-smiles-creator*) that enables a user to easily create different types of smile on a virtual character's face (Figure 1). Through radio buttons on an interface, the user could generate any smile by manipulating a combination of seven parameters: amplitude of smile, duration of the smile, mouth opening, symmetry of the lip corner, lip press, and the velocity of the onset and offset of the smile. We have considered two or three discrete values for each of these parameters (e.g., small or large for the amplitude of the smile). These parameters were selected as being pertinent in smiling behaviors (Ochs et al., 2011). When the user changes the value of one of the parameters, the corresponding video of a virtual character smiling is automatically played. Considering all the possible combinations of the discrete values of the parameters, we have created 192 different videos of smiling virtual character. We have considered three types of smiles: amused, embarrassed, and polite smiles. The user was instructed to create one animation for each type of smile. Three hundred and forty eight participants (with 195 females) with a mean age of 30 years created smiles. We then collected 348 descriptions for each smile (amused, embarrassed, and polite). In average, the participants were satisfied with the created smiles (5.28 on a Likert scale of 7 points).

2.2. Description of smiles corpus

Through the *E-smiles-creator*, we collected 1044 smile descriptions. The smiles are automatically described in terms of their types (amused, embarrassed, and polite) and their morphological characteristics. Globally, the amused smiles are mainly characterized by large amplitude, open mouth, and relaxed lips. Most of them also contain the activation of the Action Unit 6 AU6 (cheek raise), and a long global duration. Compared to the amused smiles, embarrassed smiles often have small amplitude, closed mouth, and tensed lips. They are also characterized by the absence of AU6. The polite smiles are mainly characterized by small amplitude, closed mouth,symmetry in lips shape, relaxed lips, and an absence of AU6.

2.3. Corpus-based analysis

In order to analyze the smiles corpus, we have used a *machine learning technique* called decision tree learning algorithm to identify the different morphological and dynamic characteristics of the amused, embarrassed, and polite smiles of the corpus. The input variables (predictive variables) are the morphological and dynamic characteristics and the target variables are the smile types (amused,

Figure 1: Screenshot of the E-smiles-creator.

embarrassed, or polite). Consequently, the nodes of the decision tree correspond to the smile characteristics and the leaves are the smile types. We have chosen the decision tree learning algorithm because this technique has the advantage to be well-adapted to qualitative data and to produce results that are interpretable and that can be easily implemented in a virtual agent. To create the decision tree, we took into account the level of satisfaction indicated by the user for each created smile (a level that varied between 1 and 7). More precisely, in order to give a higher weight to the smiles with a high level of satisfaction, we have done oversampling: each created smile has been duplicated n times, where n is the level of satisfaction associated to this smile. The resulting data set is composed of 5517 descriptions of smiles: 2057 amused smiles, 1675 polite smiles, and 1785 embarrassed smiles. We have used the method CART (Classification And Regression Tree) (Breiman et al., 1984) to induce the decision tree. The resulting decision tree is composed of 39 nodes and 20 leaves. All the input variables (the smile characteristics) are used to classify the smiles. With a 95% confidence interval of 1.2%: the global error rate is then in the interval [26.55%, 28.95%]. Finally, we have proposed an algorithm that enables one to determine the morphological and dynamic characteristics of the smile that a virtual agent should express given the type of smile and the importance that the user recognizes the expressed smile (value computed based on the error rate). The advantage of such a method is to consider, not only one amused, embarrassed, or polite smile but several smile types. That enables one to increase the variability of the virtual agents expressions. Compared to the literature on human smiles (Ambadar et al., 2009; Ekman et al., 1982; Keltner et al., 1995), the decision tree contains the typical amused, embarrassed, and polite smiles as reported in the literature, but it contains also amused, embarrassed, and polite smiles with other morphological and dynamic characteristics.

2.4. Validation

To validate the virtual characters smiles, an evaluation of four of the best classified amused and polite smiles have been performed *in context*. Different scenarios (of an amused, embarrassed, or polite nature) were presented in text to the user (Figure 2). Therefore, the context is represented through the scenarios. For each scenario, video clips of virtual character's different smiles were presented. We asked users to imagine the virtual character displaying the facial expression while it was in the situation presented in the scenarios. The user had to rate each of the facial expressions on its appropriateness for each given scenario. The evaluation has been conducted on the web. Seventy-five individuals participated in this evaluation (57 female) with a mean age of 32. The evaluation revealed significant results showing that the generated smiles are appropriate to their corresponding context (for more details on the experiment, see (Ochs et al., 2011)).

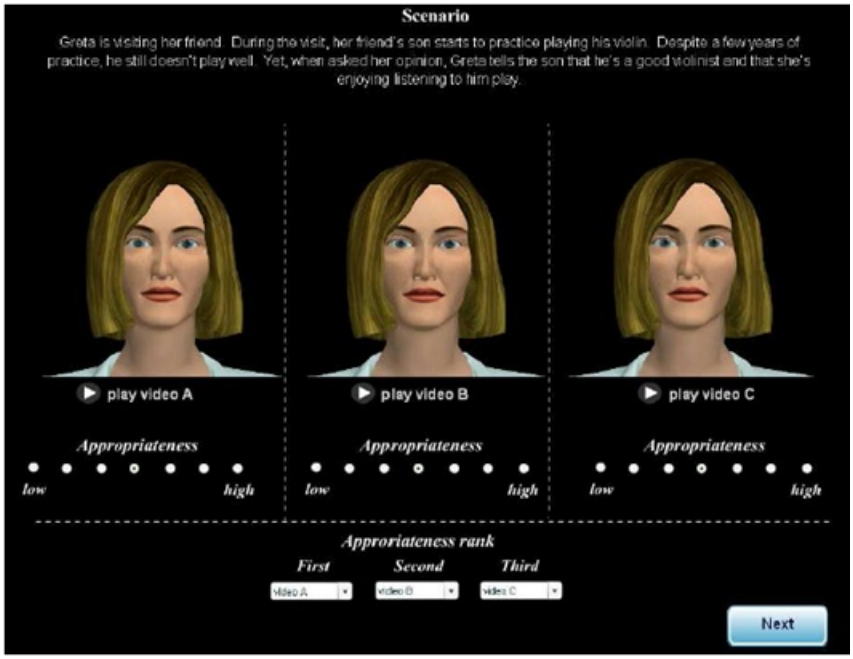

Figure 2: Screenshot of the evaluation to validate in context the virtual character's smiles.

3. Multimodal corpus of interactions between users and smiling virtual characters

3.1. Collection of data

In order to measure the effects of virtual characters smiling behavior during an interaction, we have conducted a study to collect (1) the *users behavior* (verbal and non-verbal) when interacting with a smiling virtual character, and (2) the *users perception* of a virtual character when the later displays polite and amused smiles. For this purpose, we have integrated the lexicon of smiles in the platform SEMAINE (Schröder, 2010). The smiles of the virtual character are automatically selected depending on the semantic of the virtual characters message. More precisely, the type of smile displayed by the virtual character depends on its *communicative intention*. For example, a communicative intention *encourage* or *agree* is told with a polite smile whereas a communicative intention that corresponds to the expression of*happiness* or *amusement* comes with an amused smile. In order to analyze the gender effect, two virtual characters have been considered: a female and a male (Figure 3).

Figure 3: Screenshot of the two virtual characters smiling.

The two virtual characters have exactly the same dialog behavior (i.e., same repertoire of questions and responses to the users). To measure the impact of the smiling behavior, we have implemented two versions of the virtual characters: a non-smiling version (the control condition) in which the virtual characters do not express any smile, and a smiling version in which the virtual characters display amused and polite smiles as described above. We asked to participants to interact for 3 minutes with each of two characters twice (in the two conditions). The user looked into a teleprompter, which consists of a semi-silvered screen at 45 ° to the vertical, with a horizontal computer screen below it, and a battery of cameras behind it. The user saw the face of the virtual character. The verbal and non-verbal behavior of the user was recorded through cameras and microphones.

3.2. Description of the corpus

The resulting corpus is composed of 60 audiovisual interactions between users and virtual characters (30 with the female virtual character and 30 with the male one). Each interaction lasts 3 minutes. Half corresponds to interaction with smiling virtual characters. Globally, the time of speech of the user are significantly higher than the time of speech of the virtual character since the later has more the role of a listener. Each interaction clip has been rated by the user on the following aspects: the *naturalness* of the interaction, the users *involvement* in the interaction, and the users perception of the virtual characters social stances: *polite, amused, warm, spontaneous, fun, boring*, and *cold*. These variables have been rated by the users on a scale between 0 and 10 After each interaction, we asked the user to rate his/her perception of the virtual character by answering questions (e.g. "Did you find the character was warm?").

3.3. Corpus-based analysis

Through statistical analysis, we aim at analyzing manually both the users behavior (smiles, time of speech, etc.)

and his/her self-reported perception of the virtual character to study the effect of virtual characters smiling behavior. Moreover, we will study the effect of virtual characters gender on the user: on her answers to the questionnaire and on her behavior. The results will be used to develop an algorithm that automatically determines the smiling behavior of a virtual character given the social stances it aims to express (Ochs et al., 2012).

4. Conclusion

In order to create smiling virtual characters, different corpora have been created: one focusing on the virtual characters smiles on their own and the other focusing on displays of smiling virtual characters during interactions with users. Different methods have been explored to collect such corpora. The first proposed method consists in collecting a corpus of smiling virtual characters facial expressions directly created by users. With the second method, we have collected videos of users and virtual characters interactions to investigate the effects of virtual characters smiling behavior on users perception. The corpus will be studied to analyze the influence of users gender and personality on their perception of a female and male smiling virtual character.

5. Acknowledgements

This research has been partially supported by the European Community Seventh Framework Program (FP7/2007-2013), under grant agreement no. 231287 (SSPNet).

6. References

Ambadar, Z., Cohn, J., Reed, L.: *All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous.* Journal of Nonverbal Behavior 17-34 (2009) 238252

Breiman, L., Friedman, J., Olsen, R., Stone, C.: *Classification and Regression Trees.* Chapman and Hall (1984)

Ekman, P., Friesen, W.: *Felt, false, and miserable smiles.* Journal of Nonverbal Behavior 6 (1982) 238252

Keltner, D.: *Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame.* Journal of Personality and Social Psychology 68(3) (1995) 441454

Krumhuber, E., Manstead, A., Cosker, D., Marshall, D., Rosin, P.: *Effects of dynamic attributes of smiles in human and synthetic faces: A simulated job interview setting*. Journal of Nonverbal Behavior 33 (2008) 115

M. Ochs, R. Niewiadomski, P. Brunet, and C. Pelachaud. *Smiling virtual agent in social context.* Cognitive Processing, Special Issue on Social Agents,2011.

M. Ochs, and C. Pelachaud, *Model of the Perception of Smiling Virtual Character.* Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), June 2012, Valencia, Spain.

M. Schröder. *The SEMAINE API: Towards a Standards-Based Framework for Building Emotion-Oriented Systems.* Advances in Human-Computer Interaction, 2010.

Theonas, G., Hobbs, D., Rigas, D.: *Employing virtual lecturers facial expressions in virtual educational environments.* International Journal of Virtual Reality 7(2008) 3144

Ridiculization in public debates: making fun of the other as a discrediting move

Isabella Poggi, Francesca D'Errico, Laura Vincze
Dipartimento di Scienze dell'Educazione – Università Roma Tre
Via dei Mille 23 – 00185 Roma
E-mail: poggi@uniroma3.it; fderrico@uniroma3.it, laura.vincze@gmail.com

Abstract

The paper analyzes acts of ridiculization in public debates. Ridiculization is a communicative act that, through expressing a negative evaluation of lack of power on some person, makes her feel abased, isolated, dropped out of the group, thus fulfilling a function of moralistic aggression and one of enhancing group identity. Ridiculization is seen here as a way to discredit the opponent in a political debate, in such a way as to make him/her less credible and less persuasive in front of the audience. Several cases of ridiculization are presented, and the cues to ridiculization acts are listed, from smile and laughter to simple serious words, to pretended compassion and praise, to irony, imitation and parody.

Keywords: Discredit moves. Ridiculization. Political debates.

1. Discredit in public debates

In a public debate a participant A attempts to persuade the audience that he is right and that the course of action he proposes should be taken, while his opponent B is wrong. He can do so by argumentation, that is, by illustrating his good reasons and by attacking the arguments of the other. But in some cases A attacks the opponent himself, that is, he tries to discredit him. In fact, etymologically to "dis-credit" means to make a person less credible, and, since people are persuaded not only by what one proposes but also by who proposes something, when one is not credible people are less likely to be persuaded by him.

Discredit can be defined (Poggi et al, 2011) as the spoiling of the image of a person B in front of other people C, caused, either deliberately or not, by another person A through performing communicative acts that mention or point at actions or qualities of B that are subject to a negative evaluation by the third party C.

To discredit a person one has to make a "discrediting move", that is, to pinpoint a specific feature of the person and to attribute it a negative evaluation. In a previous paper Poggi et al. (2011) analyzed the discrediting moves in political debates, that is, the communicative verbal and bodily acts by which participants try to discredit each other, and distinguished various types of them according to the "target features", i.e., the features of the Victim one points at to discredit him.

In political debates, a politician is evaluated as to three dimensions.

1. The first is Benevolence, that is, his being or not one who cares for the electors' goals; and a negative evaluation on this dimension may point at his being immoral, dishonest, or cheater.
2. On the dimension of Competence, that is, his being endowed with expertise, skill, knowledge, planning and reasoning capacity, one may cast doubts on the opponent being ignorant or stupid.
3. Finally, Dominance is the capacity of influencing others and imposing one's will. Within this dimension, a politician might be stigmatized as *helpless*, *inconsequential* or *ridiculous*.

Therefore, so far, being ridicule has been viewed as a specific target property of particular person. Yet, new observation has later convinced us that that we may that being ridiculous should be better viewed not as a feature of the discredited person, but rather as a peculiar way of stigmatizing any feature of him on any of the three dimensions: not only dominance, but also competence and benevolence.

In this work we provide a qualitative analysis of verbal and bodily discrediting moves, in TV broadcasted debates, that make use of ridiculisation.

2. Ridiculisation in public debates

Ridiculisation belongs to the category of communicative acts that convey a negative evaluation of someone. A negative evaluation (Miceli and Castelfranchi, 1998) is the belief that some object, event or person does not have (or does not provide someone with) the power to achieve some goal. One may evaluate something negatively for two reasons: either because it lacks the power to achieve some goal (negative evaluation from lack of power) or, on the contrary, because endowed with the power of thwarting some goal (negative evaluation from dangerousness).

When a person A ridicules another person B, A is conveying a negative evaluation of B for lack of power (Castelfranchi, 1988), but a lack of power that contrasts with some pretence of superiority exhibited by B. This contrast between pretence of power and actual lack of power, whose outcome is though not threatening for A, is something that can elicit laughter. In fact, laughter is a physiological expression of relief that follows the sudden disconfirmation of some expectation of danger, thus resulting in a sense of superiority (Bergson, 1900).

Thus, ridiculisation may be defined as a communicative act through which

1. a Sender S remarks, in front of some audience A, a feature or a victim V that is worth a negative evaluation of lack of power; this contrasts with V's pretence of superiority, and is seen as not threatening for S and A, so much so as to elicit relief and laughter;
2. The Sender deliberately solicits the audience A to laugh at V.

The effects of S's eliciting laughter, that are in fact all intended, i.e., wanted effects in an act of ridiculisation, are that:

a. S and A feel superior to V, because they feel above the inadequacy exhibited by the victim, and not threatened by it;
b. This common superiority strengthens the social bonds between S and A, through the shared positive emotion of laughing together, through feeling similar to each other as opposed to different from V, and through feeling a sense of alliance and complicity;
c. For the Victim, ridiculisation has the effect of abasing him/her, attacking his/her image and possibly self-image, and inducing emotions of shame and humiliation and a sense of feeling different, rejected and isolated from the group.

In a previous paper, Poggi (2010) analyzed some verbal and multimodal cases of ridiculisation in a judicial debate, where the prosecutor and the accused try to make fun of each other. In that debate during the "Mani pulite" (clean hands) trial, a trial with very important political outcomes in Italy, where many formerly powerful politicians were accused of corruption, the prosecutor ridiculises what the accused says, with the goal of lowering his credibility, convincing the judge that he is deceiving, and thus inducing him to condemn the accused; his use of ridiculisation is then a way to "dis-credit" the other in the etymological sense – making him less credible – and has a strictly judicial function. On the other hand the accused, who is a formerly very powerful politician, ridiculises the prosecutor to discredit him in the strict sense: to cast on him an image of someone who wants to exploit this trial to become famous and powerful.

This work presents a qualitative analysis of verbal and multimodal acts of ridiculisation used in political debates as a way to discredit the opponent. A semantic typology of them is outlined, and the signals through which ridiculisation is conveyed are overviewed.

3. Being ridicule and ridiculising

Ridiculising someone is in general an act performed on purpose by deliberately singling out a feature or an act of another person and pointing at it before other people as worth being laughed at. But sometimes the "Victim" makes the task easier, so to speak, to the ridiculising person, by inadvertently doing something that is "objectively" ridicule. We may call this "involuntary humor", or "making oneself ridiculous". Let us see an example from the "Clean Hands" trial.

(5) The accuser Di Pietro is trying to demonstrate that the accused, the politician Paolo Cirino Pomicino, received 5 billions Lire for political elections from the industry owner Dr. Ferruzzi. As an evidence for this Di Pietro remarks how strange is the fact (already acknowledged by Cirino Pomicino), that on the day after the elections he received Ferruzzi at his home at 7.30 in the morning. He does so to argue that Pomicino did know he was doing an illicit thing. After several turns in which Di Pietro is plying him with questions about this, Cirino Pomicino says:
CP**:** *Io credo di capire che il dottor Di Pietro voglia sostenere che io in realtà ho ricevuto il dottor Sama e il dottor Ferruzzi in quanto sostenitori finanziari della mia campagna elettorale. (...) E su questo aspetto, Presidente, io intendo rappresentare, a lei e ai suoi colleghi, che io ho ricevuto a casa mia anche persone che non hanno mai sostenuto campagne elettorali*
DP: *E ci mancherebbe, perché tutti pagano per venire da lei?*
Everybody *laughs* in the courtroom.

CP**:** I think Dr.DiPietro holds that I received Dr.Sama and Dr. Ferruzzi just because they had sustained my campaign. (…). And about this issue, Mister President, I intend to represent, to you and your colleagues, that I received at my home also people that have never sustained campaigns
DP**:** Why, does everybody pay to come to you?
Everybody *laughs* in the courtroom.

To counterargue that his receiving two industry owners at his home is not an evidence of them paying him five billions, Cirino Pomicino in a very solemn way states that he has received at his home also people that never sustained his campaign, while not realizing how funny this statement looks. Di Pietro by his rhetorical question unmasks and stresses how funny this statement looks, and everybody in the courtroom starts laughing.
Here Di Pietro is taking advantage of Cirino Pomicino inadvertently making fun of himself.

4. How to analyse cases of ridicule

In a corpus of 15 video-recorded political debates held during Italian election campaigns in 2008 and 2011, 30 fragments were selected in which a debater is mocking another (Brownell et al. 1990). For each fragment, each word, gesture, gaze, facial expression, head movement, posture was described in terms of its physical parameters (shape, movement, duration, amplitude....) and interpreted as to its literal and indirect meaning (Poggi 2007). Here is an example of how gesture and speech are coded.

1. Time	2. Speech	3. Gesture description	4. Literal meaning	5. Indir. Meaning
1 0.00.1	*"Si è detto recentemente con ironia"* *Recently people ironically said*	hands palms up oblique open outward Sp.ext: +1 Fluid: +1 Power: -1 Temp.ext: 0 Rep.: 0	Open, public I show, exhibit, show off	
2 0.00.6	*"Ma guarda Prodi fa il discorso con la CGIL e con la confindustria"* *Ok look Prodi is talking to both trade unions and factory owners*	Left arm near body, hand on Hip + Shoulder shaking Sp.ext: 0 Fluid: +1 Power: - 1 Temp.ext: 0 Rep.: 0	I am miming those who ironically judge by looking down to us	I want you to laugh about them

(1) Prodi quotes an ironic objection to his political action in order to counter-object to it.

At line 1, while saying "*Si è detto recentemente con ironia*" (recently people ironically said) (col. 2, speech), *his hands, with palms up a bit oblique, open outward* (col.

3): an iconic gesture referring to something open, public (col. 4); a way to open a new topic in your discourse, like when the curtain opens on the stage: a metadiscursive gesture, but with no indirect meaning (blank col.5). Then (line 2), while saying *"ma guarda Prodi fa il discorso con la CGIL e con la confindustria"* (Oh look, Prodi is talking to both trade unions and factory owners, col.2), *he puts his left hand on his hip*, and at the same time, with his *chest erected, he shakes his shoulders* (first left shoulder forward and right backward, then the reverse) (col.3). His *hand on hip* bears the meaning of someone taking the stance of a judge, the *erected chest* shows self-confidence, almost, a self attribution of superiority, and *shoulders shaking* shows that he is gloating for the other being judged and ridiculed.

5. Cases of ridiculisation. A semantic typology

After coding, the fragments analyzed were classified, from the semantic point of view, in terms of their target feature. As hypothesised above, it was found that a Victim may be ridiculed by stigmatizing all three target features. Let us see some examples.

a. Benevolence

The case of Prodi above can be seen as targeting the feature of benevolence. Prodi's gestures of hand on hip and shoulder shaking are a way to mimic those who said the sentence he is quoting, while making fun of them. Actually, he is somehow meta-ironizing: he is being ironic about others' irony, by ridiculing their attitude of superiority and malice towards him, through his exaggerated imitation. Irony in fact is often brought about through hyperbole, exaggeration (Attardo et al. 2003).

b. Competence

In another case ridicule is used to stigmatize the other's competence:

(2) Marco Travaglio, a left wing journalist, is talking about the numerous indictments of the right wing premier Berlusconi. Elisabetta Casellati, a vice-minister of Berlusconi's government, trying to demonstrate that not only the chief of her party has pending indictments in many trials, alludes to trials for defamation in which Travaglio has been condemned. Travaglio replies: *"...Facciamo una puntata sui miei processi, che non riguardano* […] *prostituzione minorìle, corruzione di testimòne, concussione della questura, frode fiscale per centinaia di milioni di èuro* […] *riguardano degli articoli scritti sul giornale che non sono piaciuti a qualcùno, soprattutto perché ho criticato qualcùno"*.

"Let us have a talk show about my trials, that do not concern child prostitution, witness corruption, police bribery, tax fiddle for hundreds of millions euros […]; they concern some articles written on a newspaper that someone did not like, mainly because I have criticized someone".

Travaglio utters *"articoli scritti su un giornale"* (articles written on a newspaper) while *articulating words* and *stressing tonic syllables*, and at the same time he *moves his hand, palm to Interlocutor, with joint thumb and index, rightward*, iconically depicting the action of writing. The *singsong intonation* and the clarity of depiction convey a "didactic" attitude, as if addressing to a small child, thus indirectly implying that his interlocutor (Casellati) is stupid.

c. Dominance

Finally, the opponent's dominance is targeted in the following cases:

(3) The Moderator is interviewing Margherita Hack, an old famous Italian scholar in astrophysics, who is talking against Berlusconi and the laws that he made only to save himself from trials. Roberto Formigoni, a politician on Berlusconi's side, while looking at her, shows an asymmetrical smile, with left lip corner raised, and oblique eyebrows, expressing ironic compassion.

By exhibiting compassion Formigoni implies that the opponent lacks of dominance and of the power to impose her opinion.

(4) During a talk show in a leftist TV channel, the right-wing Minister Ignazio La Russa often interrupts the left-wing Deputy Antonio Di Pietro. The Moderator Bianca Berlinguer defends Di Pietro from his interruptions and La Russa says: Ma povero, poverino (Oh poor, poor thing!), while protruding his lips as if almost crying for compassion.

La Russa too, just like Formigoni above, exhibits ironic compassion, by implying Di Pietro's lack of dominance. Beside insinuating (seriously, not kidding) that Berlinguer is not impartial because she defends a speaker of her own political side, he implies (through irony) that Di Pietro needs to be defended by her.

6. How to make fun of the opponent: signals of ridiculisation

After distinguishing cases of ridiculisation from the semantic point of view, as targeting different target features of the victim, let us see this type of communicative act from the point of view of the signals that reveal the intent of mocking the other.

Two first obvious signals are laughter and smile.

Laughter is the most typical signal conveying ridiculisation. A laughter, especially with its movements of head tilted back, shows the relaxation of someone who does not feel worried at all by the other's lack or fault; moreover, with its contagious power, it automatically calls for socialization of this feeling of relaxation and good mood, thus joining the laughing people together and making the one laughed at different, isolated, rejected.

Laughter + words. In some cases, the laughter simply metacommunicates a goal of ridiculization that, though, is made explicit by a concomitant verbal message.

Take this example from a debate between Ségolène Royal and Nicolas Sarkozy before the French president election (2007).

(5) Sarkozy has just said that, if elected, he will adopt a law to allow the mothers of handicapped children to arraign the

nursery schools that do not accept their children.
Royal *interrupts* him, she *laughs* at his proposal, and says: *Je voudrais juste dire aux femmes qu'elles n'auront pas besoin d'aller devant les tribunaux, quelle drôle de société, mais qu'elles auront le service public de la petite enfance sous toutes ses formes […] Quand les gens vont dans les tribunaux ils sont déjà débordés, quand ils ont bien d'autres choses à faire*.
Sarkozy : *Alors, je prends un autre exemple*.
Royal : *C'est pas sérieux M. Sarkozy*.
Sarkozy : *Très bien*.
Royal : *C'est pas sérieux*.
(Royal : I would like to tell women that they won't need to go to court, what a ridiculous society this would be, but they will have the public service of nursery school in all its forms. […]. When people go to court they are already overwhelmed, and they have better things to do.
Sarkozy: Then I take another example.
Royal: This is not serious, Mr. Sarkozy.
Sarkozy: Very good.
Royal: This is not serious).

While saying that people have better things to do than going to court, Royal *looks at Sarkozy* and *rapidly shakes shoulders*, as if emphasizing how pointless and absurd his proposals are. Then she overtly *laughs*, thus communicating how ridiculous is Sarkozy in making such a proposal. Still *laughing*, Royal *comfortably leans back on her chair* and *slightly rotates to right and left*, thus communicating relaxation, as if she were enjoying a comical show: so she indirectly implies that Sarkozy's proposal should not be taken seriously by the audience. But beside being implied by body communication, this is also explicitly conveyed by the word *quelle drôle de société* (what a ridiculous society), where "ridiculous" is clearly attributed to Sarkozy, and by the final sentence *C'est pas sérieux* (this is not serious).

Laughter only. In other cases, the very fact that one affords laughing – while one should be polite and respectful – counts itself as a ridiculization.

(6) Before the Referendum of 25 March 2002 pro or against immigrants' free circulation, Léonard Bender and Oskar Freysinger debate on whether Polish immigrants should be allowed to freely circulate in Switzerland or not. Freysinger, contrary to the free circulation, emphasizes the disadvantages for both Swiss and Polish people in adopting such a law.
Freysinger: *On est en train d'extraire de ces pays l'énergie vive, les personnes qui sont les plus dynamiques, les personnes qui pourraient reconstruire le pays, qui pourraient être une valeur ajoutée pour le pays et on les fait venir comme esclaves…*
Bender : *C'est pas vérifié sur le terrain.*
Freysinger : *… chez nous pour travailler à des salaires qui défient toute concurrence.*
Bender : *C'est pas vérifié sur le terrain.*
Freysinger: (scoppia a ridere : risata di divertimento con scopo di ridicolizzazione + sguardo intorno a destra e a sinistra). Mais c'est clair, Monsieur.

Freysinger: We are about to remove the live energy from this country, the most dynamical people, those who could rebuild the country, who could be an added value to the country and we make them come here as slaves…
Bender: This is not verified on the field.
Freysinger: … to work here for a salary which defies all concurrence.
Bender**:** This is not verified on the field.

Freysinger looks around with wide open eyes, first right and then left and bursts into laughter while saying: *"Mais c'est clair, Monsieur"* (But it's obvious, sir). Displayed astonishment in front of the opponent's statements is also ridiculing. The very fact of being astonished, even shocked by the interlocutor's sayings indicates that we assess his sayings' as wrong, inadequate. The fact that Freysinger also chooses to laugh at his opponent's statements communicates that Bender is not only wrong, but also ridicule. Freysinger can afford to laugh at his opponent because his inadequacy is not dangerous but only ridicule. The fact that he looks around in search of allies to laugh with him is typical of the ridiculing act. In fact, laughing at is an aggressive way of delimiting social groups and marginalizing the laughed at.

Smile + words. Similarly, smile too may in some cases simply accompanies a verbal ridiculisation

(7) December 2010. The Parliament is supposed to give a Vote of Trust in favour to Silvio Berlusconi's permanence as Prime Minister. In the political talk show "Ballarò", the two guests in the studio, both previously members of the same political party ("Popolo delle Libertà" with Berlusconi as President), now belong to two different parties. Minister Sandro Bondi is still a member of the party Popolo delle Libertà, while Italo Bocchino is now a member of the party Futuro e Libertà. The two ministers talk about some parliamentary members who presumably changed their vote due to having received some personal benefits in return by Berlusconi.
Italo Bocchino provokes minister Sandro Bondi, who answers: "*Non mi interessa discutere con Lei*[1]*"* (I am not concerned in talking to you!)
Bocchino: "*Lui mi dà del Lei perché Berlusconi gli ha detto di fare così"* (He calls me "Lei" because Berlusconi has told him to do so). And he *smiles*.
Bondi: *Le do del lei perché io sono abituato a…*
(I call you "Lei" because I am used to…)
Bocchino says: *"Il dottore gli ha detto di fare così"*. (The doctor told him to do so), while exhibiting a *large smile* and *looking at Bondi*.

Bocchino's is making fun of Bondi, first by stating he only does only what his "boss" tells him to, then displaying a first large smile of satisfaction. Immediately after, Bocchino mentions the doctor, another authority to whom Bondi obeys. In both cases Bocchino's irony and ironic smiles are definitely directed to ridicule his opponent (an act of discredit on the dominance dimension) implying that all he does is to obey orders given by his party leader, even as far as

[1] In Italian "Lei" (comparable to Engl.Thou) is a form of polite formal addressing and is opposed to "Tu" (Engl. You), generally employed when informally addressing an interlocutor one is familiar with.

relations with colleagues are concerned.
In other cases, like with Formigoni (n. 3 above), the smile itself is an act of ridiculisation.

No smile and no laughter. Yet, it is important to specify that smile and laughter are neither a necessary nor a sufficient condition to state that an act of ridiculisation is being performed. According to the degree of antagonism and to the type of debate under analysis, smiles and laughter of pure amusement can be found. The aim of laughter in that case is not one of marginalizing or excluding the other, but on the contrary, of including him in a common unifying act.
On the other hand, a participant in a debate may ridicule another without either smiling or laughing.
Like in an example, again extracted from the debate on the free circulation of Polish citizens in Switzerland, in which Mr. Bender, with a perfectly serious facial expression, ironically compares his opponent's skills in dealing with numbers and statistics to Mike Tyson's artistic skills in figure skating.

(8) Bender: *Juste une petite boutade, ne la prenais pas mal. Disons que M. Freysinger, avec les chiffres, avec les statistiques, il est aussi à l'aise que Mike Tyson en patinage artistique, tout en finesse.*
(Just a little joke, don't be offended. Mr. Freysinger with numbers and statistics is so at ease as Mike Tyson in figure skating).

Very often, especially in less explicit cases, irony is meta-communicated by facial expression – ironic smiles and/or raised eyebrows – but when the contradiction is utterly self-evident, there is no need to make it explicit by face or other signals. On the contrary, like in the following example, irony can also be signalled by a too neutral face, the so-called blank face, a visual marker of *irony* or sarcasm. (Attardo et al. 2003)
There are, though, other signals that typically accompany an act of ridiculisation, and can work s a cue to it.

Tongue in cheek. Sometimes participants in a debate, after saying something that makes fun of another participant, metacommunicate their being laughing at him/her by putting their "tongue in cheek", or by other movements produced by the mouth that reveal the Sender is trying (or better, pretending to be trying) to conceal his / her smile or laughter.

Lick lips. Another recurrent signal after a verbal ridiculization is licking one's lips. This seems to be a signal of satisfaction after a blow given to one's own opponent, and it is sometimes accompanied by a small smile of revenge or by a suppressed smile; in this last case it is sometimes present a mouth expression consisting in turning the high lip down to suck lower lip: a sort of simulation of biting one's lips to avoid biting; that is, refraining from an aggressive act.

Look around. Another typical signal of ridiculisation is when the Sender, while or after saying something – and possibly smiling or laughing, looks around himself. This may be a request for approval to the Audience, but sometimes it is a way to allude to what one has just said while asking the Audience to laugh with onself, or anyway to agree with one's allusion to the other's lack or fault. Here is an example.

(9) The leftist Moderator Michele Santoro is giving the floor to Luigi De Magistris, a former judge, now a leftist politician, who is criticizing the law proposed by the right against wiretappings. Roberto Castelli and Niccolò Ghedini, two deputies from the right, are trying to argue that Santoro is only listening to De Magistris but not to them.
Castelli says: *Ci vuole ascoltare o no?* (Will you listen to us or not?)
Santoro *turns to* Castelli and says: *Io ascolto* (I am listening)
Castelli says: *Grazie* (Thank you), while *laughing* and *looking around.*
Ghedini: *Lui ascolta, ma fa parlare lui* (he (Santoro) listens, but he lets him (De Magistris) speak). He laughs with overt teeth and half-closed eyes; then he moves his right hand towards himself and says: *Fantastico!* ((Fantastic).

Castelli, with his rhetorical question (Will you listen to us or not?), asks Santoro to listen to him and Ghedini, but when Santoro assures he is listening, Castelli ironically thanks him, thus insinuating that he feels Santoro's answer as only a polite concession, but that Santoro is I fact not impartial. Ghedini makes this insinuation explicit by saying that Santoro says he listens, but only wants to listen to De Magistris, and then ironically commenting this is fantastic. The ironic "thanks" by Castelli is accompanied by him laughing and turning his head around.: this is partly a request to the Audience to agree but also to laugh with himself to stigmatise Santoro's unfair conduction.

Imitation and parody. Other cues to the presence of ridicule are not specific signals produced by its Sender, but rather particular aspects of the Sender's communicative act, among which the presence in it of exaggeration and of imitation.
An example of both is the case of Prodi above (ex. n.1). Prodi, to make fun of those from the right that tried to make fun of him by looking down to him, is now imitating their attitude of superiority by exhibiting a haughty posture himself: thus he is imitating them but not in a faithful way, rather as in a way that resembles a parody, a caricature. Actually, a parody and a caricature are two types of communicative act – a verbal and a graphic one – that aim at causing amusement or laughter about some object (a person, a poem, a song), and do so by singling out the most characterizing features of that object and by exaggerating them in a negative way, that is by highlighting its negative aspects.

Irony. Other ways to ridicule are exploiting irony in various kinds of communicative act: for example, pretended compassion, pretended admiration, or pretended praise.
Two cases of pretended compassion are the ones seen above of Formigoni (ex. 3) and LaRussa (ex. 4).
Here is a case of pretended praise from the Swiss debate on Polish immigrants.

Mr. Bender, in favor of the free circulation of Polish, highlights a contradiction in his opponent's speech, which he calls "phenomenal".
Bender: *Et puis il y a une contradiction que je trouve assez phénoménale dans votre argumentation*. (…) *Donc un peu plus de coherence.*
(And then there is a contradiction which I consider quite phenomenal. (…) So, a little bit more coherence).

The irony of this praise is signaled by the opposite valence of the two terms: "contradiction" (negative valence) and "phénomenale" (positive valence). In fact, it is not a praise, but a pretended one, and what is required from the opponent Freysinger, as Bender concludes at the end of his turn, is "a bit more coherence".

7. Conclusion

Ridiculization is a communicative act aimed at highlighting some flaws of a person that are not, though, worrying, nor threatening; therefore to cast ridicule on someone is like telling him that he has no power over us, neither the power of making us worry. So laughing at someone – showing that particular relief we feel as we realize nothing threatens us – is a way to feel another impotent, abased, and at the same time to point at that "different" person is a way to heighten complicity with our group. If all of this is done in a public debate, as a persuasive strategy, it counts as a discrediting move: a way to convey that the other is helpless, inconsequential, to lower his credibility and his persuasive power. In this paper we have analyzed some cases of ridiculization and found out the signals that convey it or accompany it. While so far we have simply provided a qualitative analysis of single cases, in future work more extensive quantitative studies will be carried out, to investigate differences in gender and culture in the use of ridiculization and in the reactions to it.

Acknowledgements

Research supported by the Seventh Framework Program, European Network of Excellence SSPNet (Social Signal Processing Network), Grant Agreement N.231287.

References

1. Attardo S. (1994) Linguistic Theories of Humor. The Hague-Berlin: Mouton-de Gruyter.
2. Attardo, S. , Eisterhold, J., Hay, J., Poggi, I. (2003). Multimodal markers of irony and sarcasm. In Humor International Journal of Humour Research. Vol 16, 2. Walter de Gruyter GmbH & Co. KGBerlin, Germany, 243-260.
3. Bergson, H. (1900) Le rire. Essai sur la signification du comique. Paris: Éditions Alcan.
4. Brownell, H. & Gardner, H. (1988). Neurological insights into humor. In J. Durant and J. Miller (Eds.), *Laughing Matters.* London: Longman.
5. Castelfranchi C. (1988). Che figura. Emozioni e immagine sociale. Bologna: Il Mulino.
6. Draitser E.A. (1994) *Techniques of Satire: The Case of Saltykov-Scedrin*. The Hague-Berlin: Mouton-de Gruyter.
7. Miceli, M., Castelfranchi, C. (1998). The role of evaluation in cognition and social interaction, In: Dautenhahn, K. (ed.), Human cognition and agent technology. Amsterdam: John Benjamins.
8. Poggi, I. (2007) Mind, Hands, Face and Body. Weidler-Verlag Publishing House.
9. Poggi, I. (2010) Irony, humour ad ridicule. Power, image and judicial rhetoric in an Italian political trial. Publications de LÚniversité de Provence, 2010
10. Poggi, I., D'Errico, F., Vincze, L. (2011a) Discrediting moves in political debate. In F.Ricci et al. (eds) Proceedings of Second International Workshop on User Models for Motivational Systems: the affective and the rational routes to persuasion (UMMS 2011) (Girona) Springer LNCS.pp. 84-99 (ISSN 1613-0073).
11. Poggi, I., D'Errico, F. (2011b) Discrediting signals. A model of social evaluation to study discrediting moves in political desbated. In Special Issue of Journal on Multimodal User Interfaces.
12. Ruch, W. (ed) (1998). *The Sense of Humor: Explorations of a Personality Characteristic.* The

Estonian Emotional Speech Corpus: theoretical base and implementation

Rene Altrov, Hille Pajupuu
Institute of the Estonian Language
Roosikrantsi 6, 10119 Tallinn, Estonia
E-mail: rene.altrov@eki.ee, hille.pajupuu@eki.ee

Abstract

The establishment of the Estonian Emotional Speech Corpus (EESC) began in 2006 within the framework of the National Programme for Estonian Language Technology at the Institute of the Estonian Language. The corpus contains 1,234 Estonian sentences that express anger, joy and sadness, or are neutral. The sentences come from text passages read out by non-professionals who were not given any explicit indication of the target emotion. It was assumed that the content of the text would elicit an emotion in the reader and that this would be expressed in their voice. This avoids the exaggerations of acted speech. The emotion of each sentence in the corpus was then determined by listening tests. The corpus is publicly available at http://peeter.eki.ee:5000/.
This article gives an overview of the theoretical starting-points of the corpus and their usefulness for its implementation.
Keywords: emotional speech corpus, elicited emotions, non-acted speech, perception of emotions

1. Introduction

The Estonian Emotional Speech Corpus (EESC) is the only publicly available corpus containing samples of Estonian emotional speech. The main purpose of the corpus is to serve research of emotion and language technology applications (see http://peeter.eki.ee:5000/).
The creation of the corpus began by formulating theoretical starting-points (Altrov, 2008), based on overviews of existing emotion corpora and previous emotion research (Campbell, 2000; Cowie & Cornelius, 2003; Douglas-Cowie et al., 2003; Scherer et al., 2001; Ververidis & Kotropoulos, 2006). Several questions concerning the scope of the corpus and data selection had to be answered: 1) Which emotions should the corpus cover? 2) Should the corpus contain spontaneous, elicited, or acted emotions? 3) Should the texts in the corpus be spoken, or read? 4) Which texts should be selected and of what length, content and context? 5) Should the texts be presented by professional, or trained speakers (actors, announcers), or non-professionals (ordinary people)? 6) What size should the corpus be? 7) How many readers/speakers should be used? 8) Whom and how many people should be used as emotion evaluators in the perception tests?

2. Theoretical starting-points and creation of the corpus

The main decisions taken concerning the establishment of the corpus were (Figure 1):
1) Initially three emotions: sadness, anger and joy, plus neutral speech were included in the corpus as being the most useful emotions for language technology applications (Campbell, 2000; Iida et al., 2003). In this corpus these three emotions also include other related similar emotions. Thus, joy includes gratitude, happiness, pleasantness and exhilaration present in the reader's voice; sadness includes loneliness, disconsolation, concern and hopelessness; and anger includes resentment, irony, reluctance, contempt, malice and rage. Neutral speech in the corpus is normal speech without any significant emotion.
2) Simulated emotions and actors were not used due to concerns that actors might overact and use emotions that are too intense and prototypical, and therefore differ from speech that would be produced by a speaker experiencing a genuine emotion (Campbell, 2000; Iida et al., 2003; Scherer, 2003).
Authentic and moderately expressed emotions were to be gathered from text passages read out by non-professionals. The presumption was that the context of the text would stimulate the reader to express the emotion contained therein without them being told which emotion to use (Iida et al., 2003; Navas et al., 2004).
The text passages chosen were journalistic texts, unanimously recognised by readers in a special test, to contain the emotions of joy, anger or sadness. The reason for choosing journalistic texts was that when the corpus was created, it was primarily seen as being a tool for the text-to-speech synthesis of journalistic texts.
The person to read out the texts was chosen very carefully: they had to have good articulation, a pleasant voice and a sense of empathy. Experts were asked to evaluate their articulation. As empathic readers are better at rendering the emotions contained in a text, the candidates were asked to take the empathy test by Baron-Cohen & Wheelwright (2004). Another test was carried out to evaluate the pleasantness of the candidates' voices and listeners were asked to pick the speaker with the most pleasant voice (Altrov & Pajupuu, 2008). Finally, a female voice was chosen and 130 text passages were recorded for the corpus. The passages were segmented into sentences, which were then available to be used in the tests to determine the emotion of sentence.
The emotional sense of each corpus sentence is determined by listening tests. The creators of the corpus were not completely sure how well listeners would do trying to identify the emotions contained in non-acted speech without actually seeing the speaker. Therefore, the participants in the listening tests were carefully chosen to

increase the success rate in the identification of the emotion.

Earlier research implies that more mature listeners may recognise emotions from vocal cues better than younger ones (e.g., students), because emotion recognition is a culture-specific skill that can be acquired only with time (Toivanen et al., 2004). Thus the creators of EESC decided to use Estonians who were over 30 and had spent their lives in Estonia.

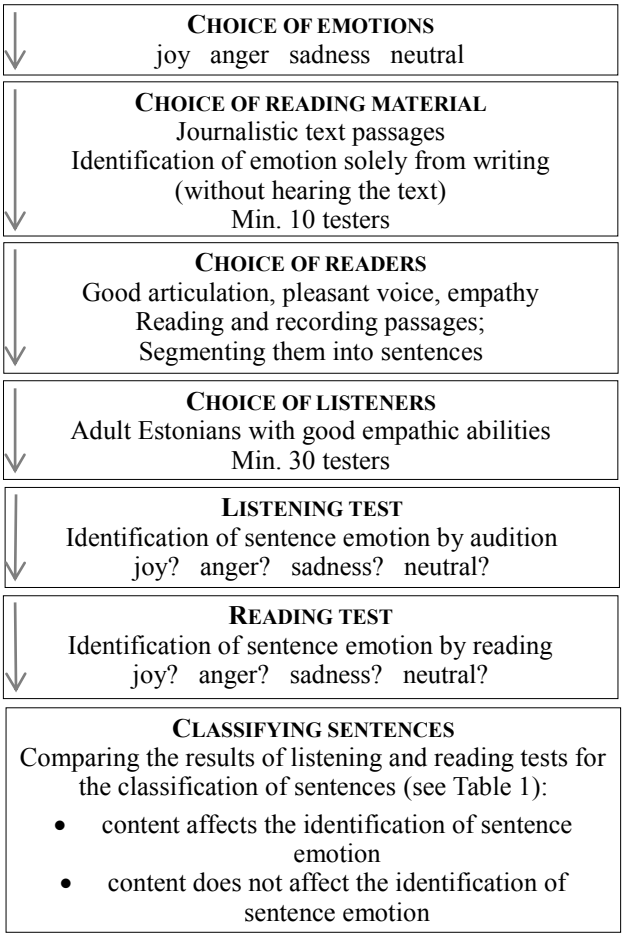

Figure 1: Creation of the EESC.

Previous studies also show that in addition to age, empathy may play a great role in the recognition of emotion (Baron-Cohen & Wheelwright, 2004; Chakrabarti et al., 2006). Relying on the presumption that empathic people are more capable of recognising emotions in voice than non-empathic people (Keen, 2006), candidates were asked to take the empathy test by Baron-Cohen & Wheelwright (2004).

Candidates were also asked, on a voluntary basis, to answer the EPIP-NEO questionnaire (for the Estonian version of the questionnaire see Mõttus et al., 2006) to study links between a person's personality traits and their ability to identify emotions.

The corpus contains 190 registered testers. Collected user data includes: sex, age, education, nationality, mother tongue, language of education, work experience, empathy quotient, and personality profile.

4) The 1,234 sentences in the corpus were used for 14 web-based tests. The underlying principle of the tests was that the content of two successive sentences must not form a logical sequence. Listening test subjects heard isolated sentences without seeing the text and then had to decide which emotion the sentences contained. The available choices were the three emotions: sadness, anger, or joy, or neutral speech.

At least 30 Estonians listened to each sentence.

In 908 sentences more than 50% of listeners identified one and the same emotion, or neutrality.

One issue with the listening tests that needed to be addressed was the role of the content in identifying the emotion of the sentence. Thus, the same sentences were used in 14 reading tests and subjects were asked to decide on the emotion or neutrality of the sentences by reading them (without audio). These subjects were not participants in listening tests.

The emotions identified by the listeners and readers did not always coincide. This led to the establishment of two categories (Table 1):

- sentences where content did not affect emotion identification (the results of reading tests differ from the results of listening tests);
- sentences where content might have affected emotion identification (the results of reading tests coincide with the results of listening tests).

Tests	Joy	Anger	Sadness	Neutral	Not sure	Sentence type in corpus
1. Ehkki Ott minu olemasolust midagi ei teadnud. [Although Ott knew nothing of my existence.]						
By listening	**87.5**	0.0	0.0	12.5	-	**Joy,** no content influence
By reading	4.0	0.0	32.0	32.0	32.0	
2. Ükskõik, mida ma teen, ikka pole ta rahul! [Whatever I do, he is never satisfied!]						
By listening	0.0	14.3	**80.0**	5.7	-	**Sadness**, no content influence
By reading	0.0	**64.3**	35.7	0.0	0.0	
Täiesti mõistetamatu! [Completely incomprehensible!]						
By listening	0.0	**100.0**	0.0	0.0	-	**Anger**, content influence
By reading	0.0	**83.0**	0.0	11.0	5.6	

Table 1: Classification of emotions in the corpus by emotion identification in reading and listening tests (test results in %).

In Table 2 the number of corpus sentences is given by groups.

Emotion	Sentences	Content influence on identification	No content influence on identification
joy	232	163	69
anger	277	177	100
sadness	191	88	103
neutral	208	87	121
unable to identify	326		
Total	1234		

Table 2: Number of sentences in emotion corpus.

Although such double testing of each Corpus sentence is rather time-consuming, it works as a validator for the

corpus. Corpus users can be sure that corpus sentences contain emotions that can be identified during listening. Users can select sentences where emotion is rendered by voice only or sentences where emotion is also rendered by content.

5) The corpus was designed so that it could be used for multiple purposes and extended by adding readers, sentences and emotions.

3. Options for corpus users

Users can search for sentences expressing anger, joy, or sadness, or neutral sentences from the corpus (http://peeter.eki.ee:5000/reports/list/).

Sentences are displayed as text and can be listened to by clicking on them. The identification rate of emotion in each sentence is also displayed.

Queries can be narrowed down to include only sentences in which:

- content did not affect the identification of emotion;
- content might have affected the identification of emotion.

The audio-recordings and text of sentences can be downloaded and saved (wav, textgrid). There are three labelling levels: phonemes, words and pauses, sentences.

4. Implementation details

The corpus is a web-based application that uses freeware: Linux, PostgreSQL, Python, Praat. All data except for audio files have been saved in a PostgreSQL database. The web interface was created and all data processing carried out by using the Python programming language and Pylons web framework. The application can be installed in Windows and Linux systems. The web interface is available for Estonian, English, Finnish, Latvian, Russian and Italian, and can be easily adapted for other languages. For the technical description of the corpus see http://peeter.eki.ee:5000/docs/

5. Preliminary results

Currently the corpus is in a stage where the validity of the theoretical starting-points can be verified and, if necessary, corrections can be made.

1. It has been confirmed that listeners can easily identify moderately expressed emotions from the voice of a non-professional reader. For 73.5% of corpus sentences over 50% of listeners identified one and the same emotion, or decided that the sentence was neutral (Altrov & Pajupuu, forthcoming), see Table 3.

Listening response	Joy	Anger	Sadness	Neutral
Emotional sentences identified by more than 50% of listeners	232	277	191	208
Mean percentage of identification and std	75.4 14.5	73.3 14.6	72.1 14.7	68.3 11.9

Table 3: Statistics of the emotional and neutral sentences identified by the listening test.

2. In the early stages of creating the EESC the decision was made to use people older than 30 as emotion identifiers. This decision relied on the assumption that people who have lived longer in a certain culture are more likely to have acquired the skills of culture-specific expression of emotions. In order to find out if the decision to use older people as corpus testers was justified, Altrov and Pajupuu (2010) compared the results of emotion identification by people older than 30 and younger than 28 and found that the two groups differed significantly. Younger people identified more sentences as neutral. Both groups were also compared with Latvians. The latter identified emotions quite differently from Estonians. From these results it can be said that the identification of emotions really is culture-specific and accurate emotion identification requires spending a longer period in a particular culture. It is therefore wise to use people who have lived in Estonia longer for identifying emotions from vocal expression.

3. Currently a study is being undertaken on how much listeners' empathic abilities affect their ability to identify emotions from vocal expression.

4. Another issue that needs to be addressed is whether classifying corpus sentences according to the influence of sentence content on emotion identification is justified, i.e., if any significant differences can be found between the acoustic parameters of the two groups – "content affects identification" and "content does not affect identification". So far, the corpus material has only been used for studying the difference in intensity of sentence emotions in the two groups. ANOVA analysis has shown that the intensity of sentences expressing anger and joy and neutral sentences in the two groups differ significantly. However, there is no such difference in intensity in sentences expressing sadness (Table 4). Although it is just one acoustic characteristic, it may mean that the content of text affects how an emotion is acoustically expressed, which also means that dividing corpus sentences into two groups is justified.

Pairs: content influences – no content influences	Df	Sum Sq	Mean Sq	F value	Pr(>F)
joy	1	103.00	103.00	5.62	0.0178
Residuals	4189	76707.60	18.31		
anger	1	271.38	271.38	11.92	0.0006
Residuals	5053	114992.73	22.76		
sadness	1	2.80	2.80	0.13	0.7166
Residuals	3467	73757.47	21.27		
neutral	1	591.40	591.40	31.66	0.0000
Residuals	3949	73755.52	18.68		

Table 4: ANOVA results on emotional intensity of sentences in two groups: "content affects identification" and "content does not affect identification".

6. Conclusion

This paper gives an overview of the theoretical base, creation and content of the Estonian Emotional Speech Corpus. The EESC contains 1,234 Estonian sentences that have passed both reading and listening tests. Test takers identified 908 sentences that expressed anger, joy,

sadness, or were neutral. The sentences were divided into two groups: sentences in which content affected the identification of the emotion and sentences in which it did not. Development of the corpus continues. Corpus sentences have also been categorised as positive, negative and neutral. Preparations for extending the corpus by adding video clips with spontaneous speech and their testing are under way. The corpus is freely available and used in the language technological projects for emotional speech synthesis, as well as for recognition of emotions.

7. Acknowledgements

The study was supported by the National Programme for Estonian Language Technology and the project SF0050023s09 "Modelling intermodular phenomena in Estonian".

8. References

Altrov, R. (2008). Eesti emotsionaalse kõne korpus: teoreetilised toetuspunktid. *Keel ja Kirjandus*, 4, pp. 261–271.

Altrov, R., Pajupuu, H. (2008). The Estonian Emotional Speech Corpus: release 1. In F. Cermák, R. Marcinkevicienè, E. Rimkutè & J. Zabarskaitè, *The Third Baltic Conference on Human Language Technologies.* Vilnius: Vytauto Didžiojo Universitetas, Lietuviu kalbos institutas, pp. 9–15.

Altrov, R., Pajupuu, H. (2010). Estonian Emotional Speech Corpus: Culture and age in selecting corpus testers. In I. Skadiņa, A. Vasiļjevs (Eds.), *Human Language Technologies – The Baltic Perspective – Proceedings of the Fourth International Conference Baltic HLT 2010.* Amsterdam: IOS Press, pp. 25–32.

Altrov, R., Pajupuu, H. (forthcoming). Estonian Emotional Speech Corpus: Content and options. In G. Diani, J. Bamford, S. Cavalieri (Eds.). *Variation and Change in Spoken and Written Discourse: Perspectives from Corpus Linguistis.* Amsterdam: John Benjamins.

Baron-Cohen, S., Wheelwright, S. (2004). The Empathy Quotient: An investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders,* 34(2), pp. 163–175.

Campbell, N. (2000). Databases of emotional speech. In R. Cowie, E. Douglas-Cowie, & M. Schröder (Eds.), *ISCA Workshop on Speech and Emotions.* Newcastle: North Ireland, pp. 34–38.

Chakrabarti, B., Bullmore, E., Baron-Cohen, S. (2006). Empathizing with basic emotions: common and discrete neural substrates. *Social neuroscience,* 1(3-4), pp. 364–384.

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

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

Iida, A., Campbell, N., Higuchi, F., Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication,* 40(1-2), pp. 161–187.

Keen, S. (2006). A theory of narrative empathy. *NARRATIVE*, 14(3), pp. 207–236.

Mõttus, R., Pullmann, H., Allik, J. (2006). Toward more readable Big Five personality inventories. *European Journal of Psychological Assessment,* 22(3), pp. 149–157.

Navas, E., Castelruiz, A., Luengo, I., Sanchez, J., Hernaez, I. (2004). Design and recording an audiovisual database of emotional speech in Basque. In International conference on language resources and evaluation (LREC), Lisbon Portugal, pp. 1387–1390.

Scherer, K.R., Banse, R., Wallbott, H.G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-Cultural Psychology,* 32(1), pp. 76–92.

Toivanen, J., Väyrynen, E., Seppänen, T. (2004). Automatic discrimination of emotion from spoken Finnish. *Language & Speech,* 47(4), pp. 383–412.

Ververidis, D., Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication*, 48(9), pp. 1162–1181.

Building Resources for Multilingual Affect Analysis – A Case Study on Hindi, Bengali and Telugu

Dipankar Das, Soujanya Poria, Chandra Mohan Dasari, Sivaji Bandyopadhyay
Department of Computer Science & Engineering, Jadavpur University
188, Raja S.C. Mullick Road, Kolkata 700 032, India
dipankar.dipnil2005@gmail.com, soujanya.poria@gmail.com, chanduthecm@gmail.com, sivaji_cse_ju@yahoo.com

Abstract

The rapid growth of affective texts in the Web 2.0 and multilingualism in search engines motivate us to prepare the emotion/affect data for three Indian languages (Hindi, Bengali and Telugu). This paper reports the development of the *WordNet Affects* and *SemEval 2007* affect sensing corpora in three target Indian languages from the available English sources that were provided in the Affective Text shared task on the *SemEval 2007* workshop. The linguistic evaluation on the developed resources proposed various morals from the perspective of affect analysis in the target languages. Two emotion analysis systems, baseline systems followed by morphology driven systems have been developed and the evaluation results of the systems produce satisfactory results in comparison with the English and Japanese.

1. Introduction

Affect analysis is a natural language processing (NLP) technique for recognizing the emotive aspect of text whereas the same textual content can be presented with different emotional slants (Esuli and Sebastiani, 2006). The majority of subjective analysis methods that are related to opinion, emotion/affect or broadly sentiment is based on textual keywords spotting and therefore explores the necessity to build specific lexical resources.

A recent study shows that the non-native English speakers support the growing use of the Internet[1]. This raises the demand of linguistic resources for languages other than English. The domain of multilingual sentiment analysis consists of related work for several European languages such as Romanian/Spanish (Banea et al., 2008; Mihalcea et al., 2007), German (Denecke et al., 2008) etc. In recent times, the sentiment-labeled data is also gradually becoming available for languages other than English (e.g. Seki et al., 2008; Wan, 2009; Prettenhofer and Stein, 2010; Nakagawa et al., 2010; Schulz et al., 2010). But, the multilingual resource crisis is still present from the perspective of affect/emotion or fine grained sentiment. In case of affect analysis, there are a few attempts in other languages, such as Russian and Romanian (Bobicev et al., 2010), Japanese (Torri et al., 2011) etc.

India is a multilingual country with great cultural diversities. But, the crucial fact is that the Indian languages are resource-constrained and the manual preparation of affect annotated data is both time consuming and cost intensive. To the best of our knowledge, at present, there is no such lexicon or corpus available for affect analysis in Indian languages except Bengali (Das and Bandyopadhyay, 2009a, 2010). Hindi, the national language of India enjoys a speaker population of 43 crore (2001 census of India[2]; rounded to the most significant digit). Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh. Recently, Telugu is receiving the attention from national and international levels[3]. The identification of affect in Indian languages in general and Hindi, Bengali and Telugu in particular is difficult and challenging as both of the languages are 1) Inflectional languages providing the richest and most challenging sets of linguistic and statistical features resulting in long and complex word forms, and 2) Relatively free phrase order. Thus, we believe that the present task would help the development and evaluation of the emotion analysis systems in other languages as well.

In the present task, we have prepared the *WordNet Affect* for three Indian languages (Hindi, Bengali and Telugu) from the already available English *WordNet Affect* (Strapparava and Valitutti, 2004). Expansion of the English *WordNet Affect*[4] synsets using *SentiWordNet 3.0*[5] has been performed to verify whether any target emotion word can be produced from a source sentiment word during translation or not. The number of entries in the expanded word lists was increased by 69.77% and 74.60% at synset and word levels, respectively. Hindi *WordNet* [6], a freely available lexical resource was developed based on the English *WordNet 3.0*[7]. Thus, the English synsets of the expanded lists are automatically translated into equivalent synsets of the Hindi language based on the *synsetID*. The development of the Bengali *WordNet Affect* was already attempted in (Das and

[1] http://www.internetworldstats.com/stats.htm

[2] http : //www:censusindia:gov:in/CensusData2001 /CensusDataOnline/Language/Statement1.htm

[3] http://sites.google.com/site/iticgift/

[4] http://www.cse.unt.edu/~rada/affectivetext/

[5] http://sentiwordnet.isti.cnr.it/

[6] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[7] http://wordnet.princeton.edu/wordnet/download/

Bandyopadhyay, 2010). Presently, we have used their resource for our affect analysis task. The *WordNet* for Telugu is being developed by different research groups. But, the resources do not perform in the required measure. Therefore, we have carried out the translation of the Telugu words in the expanded synsets using google online dictionary API [8]. The linguistic evaluation of the translated Telugu *WordNet Affects* produces a moderate agreement.

Primarily, we have prepared the baseline systems for three target languages based on the *WordNet Affects* that have been developed for the respective target languages. The baseline systems have been evaluated on the translated corpora of the target languages. The English *SemEval 2007* affect sensing corpus in (Strapparava and Mihalcea, 2007) was translated in three different target languages to serve the purpose. The translated corpus in Hindi have been prepared using Google translator API [9] followed by linguistic editing whereas the translation for Bengali and Telugu has been performed manually by the authors. The baseline systems achieve the average *F-scores* in the range from 46.39% to 56.88% with respect to six emotion classes. We have also incorporated the morphological knowledge of the emotion words into the baseline systems and the performance of the systems was increased satisfactorily.

We have compared our present results on three Indian languages with similar research results for English (Das and Bandyopadhay, 2009b) and Japanese (Torri et al., 2011). It has been observed that a perfect sense to sense mapping among languages is impossible, as a particular sense may denote additional meanings and uses in one language compared to another, thus rendering a perfect parallel sense boundary permeable (Banea et al., 2011). But, the emotional senses do hold across languages, implying that this information could be leveraged in an automatic fashion to provide additional clues for the affect labeling of unseen senses.

The rest of the paper is organized as follows. The related tasks on multilingual affect analysis and lexicon development are mentioned in Section 2. Different developmental phases of the *WordNet Affects* are described in Section 3. Preparation of translated corpora is discussed in Section 4. Different experiments and evaluations based on morphology and the annotated emotion scores are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Related Work

A growing demand of multilingual sentiment analysis provides the opportunities for building resources from one language to other languages. The pioneering works on multilingual subjectivity and sentiment analysis were attempted either by translating or leveraging a bilingual dictionary or a parallel corpus based on English resources (Mihalcea et al., 2007; Bautin et al., 2008; Banea et al., 2010). Some interesting approaches are also mentioned in (Boyd-Graber and Resnik, 2010; Ahmad et al. 2006). The English annotation schemes developed for opinionated text was carried out by (Esuli et al., 2008) when annotating expressions of private state in Dutch (Maks and Vossen, 2010). The English corpora manually annotated for subjectivity or sentiment such as MPQA (Wiebe et al., 2005), or the multi-domain sentiment classification corpus (Blitzer et al., 2007) were subjected to experiments in Spanish, Romanian (Banea et al., 2008), German (Denecke et al., 2008), Chinese (Wan, 2009) etc. In recent times, the sentiment-labeled data is gradually becoming available for languages other than English (e.g. Seki et al., 2008; Das and Bandyopadhyay, 2009; Prettenhofer and Stein, 2010; Nakagawa et al., 2010; Schulz et al., 2010). Some of the Sentiment and subjectivity lexicons were also transferred into Chinese (Ku et al., 2006) and into Romanian (Mihalcea et al., 2007).

To the best of our knowledge, at present, no efforts are found for preparing the affect related dictionaries or corpora in Indian languages. Thus, we believe that this research effort would help the development and evaluation of emotion or affect analysis systems in other languages as well.

3. Development of WordNet Affects

In the present task, we have prepared the *WordNet Affect* for three Indian languages (Hindi, Bengali, and Telugu) from the already available English *WordNet Affect* (Strapparava and Valitutti, 2004). The English *WordNet Affect* is a small lexical resource compared to the complete *WordNet* (Miller, 1995) but its affective annotation helps in emotion analysis. The entries in the English *WordNet Affect* are annotated using Ekman's (1993) six emotional categories (*joy, fear, anger, sadness, disgust, surprise*). The collection of the English *WordNet Affect* [10] synsets that are used in the present work was provided as a resource in the "*Affective Text*" shared task of *SemEval-2007* Workshop (Strapparava and Mihalcea, 2007). The shared task was focused on text annotation by affective tags. We have not considered the problems of the lexical affect representation or discussed the differences between emotions, cognitive states and affects in developing the *WordNet Affects* in the target languages. The whole data is provided in six emotions. Each file contains a list of synsets and one synset per line. An example synset entry from the *WordNet Affect* is as follows.

[8] http://www.google.com/dictionary
[9] http://translate.google.com/#

[10] http://www.cse.unt.edu/~rada/affectivetext/

a#00117872 angered enraged furious infuriated maddened

The first letter of each line indicates the part of speech (POS) and is followed by the *affectID*. The representation was simple and easy for further processing. We have mapped the synsetID of the *WordNet Affect* lists with the synsetID of the WordNet 3.0 using an open source tool[11]. This mapping helps in expanding the *WordNet Affect* lists with the recent version of *SentiWordNet 3.0*[12].

3.1 Expansion by SentiWordNet

It was observed that the six *WordNet Affect* lists that were provided in the shared task contain only 612 synsets in total with 1536 words. The words in each of the six emotion lists have been observed to be not more than 37.2% of the words present in the corresponding *SentiWordNet* synsets. Hence, these six lists are expanded with the synsets retrieved from the English *SentiWordNet* (Baccianella et al., 2010) to have an adequate number of emotion related word entries. We assumed that the new sentiment bearing words in English *SentiWordNet* might have some emotional connotation in the target languages. But, the POS information for each of the synsets is kept unchanged during expansion of the lists. The numbers of entries in the expanded word lists are increased by 69.77% and 74.60% at synset and word levels respectively.

The *SentiWordNet* assigns each synset of *WordNet* with two coarse grained subjective scores such as *positive*, *negative* along with an *objective* score. Moreover, the *SentiWordNet* contains more number of coarse grained emotional words than the *WordNet Affect*. We assumed that the translation of the coarse grained emotional words into target languages might contain more or less fine-grained emotion words. The differences between emotions, cognitive states and affects were not analyzed during translation. Our main focus in the task was to develop the equivalent resources in target languages for analyzing emotions.

As the *WordNet Affect* and *SentiWordNet* were both developed from the *WordNet* (Miller, 1995), each word of the *WordNet Affect* has easily been replaced by the equivalent synsets retrieved from the *SentiWordNet* if the synset contains that emotion word (Das and Bndyopadhyay, 2010; Torri et al., 2011). In case of word ambiguity during the replacement of the words in the *WordNet Affect* synsets, some spurious senses appeared in some synsets that represent a non appropriate meaning. But, it was observed that in the case of emotion words, this phenomenon is not frequent because the direct emotion words are not very ambiguous.

[11] http://nlp.lsi.upc.edu/web/index.php?option=com_content&task=view&id=21&Itemid=59

[12] http://sentiwordnet.isti.cnr.it/

3.2 Translation

We have used English as a source language. The translation into the target languages has been performed in different ways. The HIndi *WordNet* [13] is freely available and it was developed based on the English *WordNet*. The synsets of the expanded lists were thus automatically translated into Hindi equivalent synsets based on the *synsetIDs*. The following are some translated samples that contain word level as well as synset level translations.

a#*admirable* →
#प्रशंसनीय (*prasangsaniyo*) (2456 - ADJECTIVE - [प्रशंसनीय, प्रशंस्य, श्लाघ्य, श्लाघनीय, सराहनीय, स्तुत्य, धन्य……

#सराहनीय (*sarahaniyo*) (2456 - ADJECTIVE - [प्रशंसनीय, प्रशंस्य, श्लाघ्य, श्लाघनीय, सराहनीय, स्तुत्य, धन्य….

#श्लाघ्य (*slaghya*) (2360 - ADJECTIVE - [उत्तम, उत्कृष्ट, बेहतरीन, आला, अकरा, अनमोल, श्रेष्ठ, उम्दा, उमदा….

#अत्युत्तम (*atyuttam*) (7438 - ADJECTIVE - [अत्युत्तम])

#अनुपम (*anupam*) (2290 - ADJECTIVE - [अनुपम, अतुलनीय, अद्वितीय, अनोखा, असाधारण, लाजवाब, बेजोड़, बेमिसाल…

Some synsets (e.g., 00115193-a *huffy, mad, sore*) were not translated into Hindi as there is no equivalent synset entries in *WordNet* for those affect synsets. But, the translated synsets contain multi-word elements (कुपित होना (*kupit hona*) '*being angered*' etc.).

To the best of our knowledge, the Bengali *WordNet* is not yet freely available. Therefore, the expanded *WordNet Affect* lists were translated into Bengali using the synset based English to Bengali bilingual dictionary (Das and Bandyopadhyay, 2010). The dictionary contains 1,02,119 synsets that were developed using Samsad Bengali to English bilingual dictionary [14] as part of the EILMT [15] project. The synset-based dictionary is developed from the general domain. We have not considered all the word combinations, as they could not be translated automatically. The sense disambiguation task was conducted based on the hints of sense wise separated word groups present in Bengali to English bilingual dictionary (Das and Bandyopadhyay, 2010).

a#*admirable* → প্রশংসনীয়, অপূর্ব

[13] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[14] http://home.uchicago.edu/~cbs2/banglainstruction.html

[15] English to Indian Languages Machine Translation (EILMT) is a TDIL project undertaken by the consortium of different premier institutes and sponsored by MCIT, Govt. of India.

There is no Telugu *WordNet* that is freely available and it is being developed based on the English *WordNet*. Thus the English synsets of the expanded lists have been automatically translated into Telugu equivalent synsets based on the *synsetIDs* using open source google dictionary API. The lists were verified by the authors. There are some translated samples that contain word level as well as phrase level translations.

n#cheer → ప్రోత్సహించేందుకై కొట్టే చప్పట్లు (*prostyahinchenduky kotte chappatlu*)

v#chafe → గట్టిగా రుద్దడం వల్ల బాధ కలుగు (*gattiga rudhatam valla badha kalugu*)

The number of translated synsets (S) and words (W) for six affect lists of the three target languages are shown in Table 1.

Affect Lists	Hindi	Bengali	Telugu
anger	210 (S), 1367 (W)	321 (S) 1141 (W)	240 (S) 1033(W)
disgust	123 (S) 873 (W)	74 (S) 287 (W)	22 (S) 218 (W)
fear	235 (S) 1478 (W)	182(S) 785 (W)	80 (S) 615 (W)
joy	617 (S) 3921 (W)	467(S) 1644 (W)	379 (S) 2940 (W)
sadness	356 (S) 2128 (W)	220 (S) 788 (W)	133 (S) 846 (W)
surprise	102 (S) 712 (W)	125(S) 472 (W)	74 (S) 456 (W)

Table 1: Number of translated synsets (S) and words (W) for six affect lists of the three target languages

3.3 Analyzing Translation Errors

Some of the synset elements of the *WordNet Affect* lists were not automatically translated. One of the reasons may be that the bilingual dictionary is fairly modest. It has been observed that some words containing suffixes such as "*ness*", "*less*", "*ful*" as well as adverbs formed using suffix "*ly*" are unlikely to appear in dictionaries.

It was also found that the idioms, word combinations were not translated automatically. There were a large number of word combinations, collocations and idioms in the Telugu *WordNet Affect*. These parts of synsets show problems during translation and therefore manual translation is carried out for these types. There are some of the English words that were not translated into Telugu (e.g., *astonied, awestricken, dazed, dumbfound, howling, in_awe_of, marvelously, stupefy, superbly* etc). Some of the wrongly translated emotion words were removed from their corresponding lists (e.g., సంతానం (*santhanam*) '*brood*', సంఘటనా సమయము (*sangatana samayamu*) '*scene*' etc.).

Some *SentiWordNet* synsets (e.g., 00115193-a *huffy, mad, sore*) were not translated into Hindi as there are no equivalent synset entries in the Hindi *WordNet*. There are some of the English synsets that were not translated into Hindi. For example, the synset '07517292-n *lividity*' contains only one English word that was not translated into Hindi. One of the reasons of such translation problems may be that no equivalent Hindi word sense is available for such English words.

The development of the Bengali *WordNet Affect* has already been attempted (Das and Bandyopadhyay, 2010). But, the coverage of the Bengali *WordNet Affect* was not exemplified. It is found that number of emotion words have increased in all of the Bengali *WordNet Affect* lists except *joy* and *sadness*. The non-translated entries were filtered from the English synsets after the translation. As the total number of non-translated words in the six emotion lists is 210 and the figure is comprehensible for manual translation, the non-translated words were translated into Bengali by the authors.

4. Preparation of Translated Corpora

Knowledge resources can be leveraged in identifying emotion-related words in text and the lexical coverage of these resources may be limited, given the informal nature of online discourse (Aman and Szpakowicz, 2007). In general, the identification of the direct emotion words incorporates the lexicon lookup approach. Hence, we evaluated the developed *WordNet Affects* on their corresponding target language corpora that have been translated from *SemEval 2007* affect sensing corpus (Strapparava and Mihalcea, 2007).

4.1. Automatic Translation

The English *SemEval 2007* affect sensing corpus consists of news headlines only. Each of the news headlines is tagged with a valence score and scores for all the six Ekman's (1993) emotions. The six emotion scores for each sentence are in the range of 0 to 100.

In case of Hindi, we used the Google translator API [16] to translate the 250 and 1000 sentences of the trial and test sets of the *SemEval 2007* corpus. Some linguistic corrections were also made for the automatically translated Hindi corpus. But, the API is not available for Bengali or Telugu. Thus, the translated corpora for these target languages were prepared by the authors. The experiments regarding morphology and emotion scores have been conducted on the trial corpus. We have carried out different experiments on 1000 test sentences by selecting different ranges of emotion scores.

[16] http://translate.google.com/#

4.2. Manual Agreement

One of the major problems of emotion identification is the lack of appropriately annotated corpora. But, in the present task, the English *SemEval 2007* affect corpus was already annotated with the valence score (-100 to -1 for *negative*, 0 for *neutral* and 1to 100 for *positive*) and six emotion scores (in the range of 0 to 100) for each sentence. The annotation task on the translated corpora was also easy to comprehend as we assumed that the translation carries the emotional senses across languages. The annotators were simply asked to answer "*agree*" or "*disagree*" by viewing the translated sentence and its corresponding gold standard emotion and valence score.

Three annotators identified as A1, A2 and A3 were asked to carry out the annotation. The crucial fact was that each of the annotators was specialized in only two native languages (e.g., A1 knows only Hindi and Bengali, A2 knows only Hindi and Telugu and A3 knows only Bengali and Telugu). Thus, two-way agreement scheme was proposed, one with respect to the English gold standard (GS) annotation and other with respect to the native annotators specialized in the corresponding languages.

The annotation scope was restricted to sentence level. Thus, the annotators were bound within the sentence level without considering any surrounding context or discourse. We have used the standard metric, Cohen's *kappa* coefficient (κ) (Cohen, 1960) for measuring the inter-annotator agreement. *Kappa* is a statistical measure of inter-rater agreement for qualitative (categorical) items and measures the agreement between two raters who separately classify items into some mutually exclusive categories. We measured the *kappa* agreement with respect to each sentence.

The results of agreement with respect to all emotion classes are shown in Table 2. It was observed that the agreement of validating the sentential valence (≈ 0.9) shows moderate and acceptable values whereas the real disagreement occurs in case of validating the emotion scores. The validation of emotion scores for two or three closely relevant and related emotion classes causes the disagreement. It was found that the average number of emotion types is 2~3 which indicates the presence of multiple emotions in a sentence. It has to be mentioned that the agreement in identifying emotion in the sentences containing single emotion score is more than the agreement in identifying emotion in those sentences that contain multiple emotion scores with close values.

In addition to the above issues, some other interesting observations were found from the perspective of disagreement study. The *happy* emotion more frequently conflict with *surprise* emotion in the translated corpora rather than any other emotions. Consequently, the *sadness* occurs with *angry*, *disgust* and *fear* emotion types rather than *happy* or *surprise*. The reason may be that, at sentence level translation, the emotion tags with similar emotional slants occupy with more close association in comparison with the others that present with opposition. But, overall, the agreement results on translated corpora for all three target languages were satisfactory with respect to native nature of the languages as well as in comparison with English Gold standard annotation.

Annotators	**Hindi**	**Bengali**	**Telugu**
English Gold Standard (GS) Vs. Native Annotator			
GS ↔A1	**0.85**	**0.83**	X
GS ↔A2	**0.84**	X	**0.83**
GS ↔A3	X	**0.84**	**0.82**
Native Annotator Vs. Native Annotator			
A1 ↔ A2	**0.85**	X	X
A1 ↔ A3	X	**0.84**	X
A2 ↔ A3	X	X	**0.82**

Table 2: Inter-Annotator Agreement using *kappa*

5. Evaluation

Three baseline systems have been developed for three different target languages based on their corresponding *WordNet Affects*. The algorithm is that, if a word in a sentence is present in any of the *WordNet Affect* lists; the sentence is tagged with the emotion label corresponding to that affect list. But, if any word is not found in any of the six lists, each word of the sentence is passed through the morphological process to identify its root form and the root form is searched through the *WordNet Affect* lists again. If the root form is found in any of the six *WordNet Affect* lists, the sentence is tagged accordingly. Otherwise, the sentence is tagged as non-emotional or *neutral*.

The open source Hindi Stemmer[17] has been employed for identifying the root forms of the words in Hindi. Due to the scarcity of efficient stemmers in case of Bengali and Telugu, we have used the morphological information from the output of the open source Bengali and Telugu shallow parsers[18].

To evaluate the baseline systems, we have considered that each sentence is assigned a single sentential emotion tag based on the maximum emotion score out of six annotated emotion scores. The experiments regarding morphology and emotion scores were conducted on the trial corpus. The final evaluation which was carried out on 1000 test sentences produces the results shown in Table 3. The evaluation of our system is similar with the coarse-grained evaluation methodology of the *SemEval*

[17] http://www.cfilt.iitb.ac.in/wordnet/webhwn/

[18] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

2007 shared task on affective text. It has been observed that Hindi and Bengali performs satisfactorily while the Telugu baseline system fails due to the coverage of the Telugu *WordNet Affect* and effect of rich and deep morphology. The precision is high in case of Hindi and Telugu whereas the system achieves low recall for Bengali only. The results of the present baseline systems were evaluated in comparison with English (Das and Bandyopadhyay, 2009b) and Japanese (Torri et al., 2011). The results are shown in Table 4.

In addition to the coarse-grained evaluation, we also carried out different experiments by selecting different ranges of emotion scores. The corresponding experimental results are also shown in Table 5. Incorporation of morphology improves the performance of the system. On the other hand, it was observed that the performance of the system decreases by increasing the range of Emotion Scores (ES). The reason may be that the numeric distribution of the sentential instances in each of the emotion classes decreases as the range in emotion scores increases. This, in turn, decreases the performance of the system. Telugu affect lists include words as well as phrases. We deal with phrases using Telugu morphology tool to find affect words in a sentence and substitute an affect word into its original conjugated form. One of the main reasons of using a morphology tool is to analyze the conjugated form and to identify the phrases. For example, the Telugu word for the equivalent English word '*anger*' is "కోపం (*kopam*)" but there are other conjugated word forms such as "ఆగ్రహం తెప్పించింది (*agraham theppinchindhi*)" that means '*angered*' and it is used in past tense. Similarly, other conjugated form "కోపగించు కున్నారు (*kopaginchu kunnaru*)" which denotes the past participle form '*have angered*' of the original word '*anger*'. The morphological form of its passive sense is "కోపగించు కుంటారు. (*kopaginchu kuntaru*)" that means '*be angered*'. In addition to that, we identify the words into their original forms from their corresponding phrases by using the morpheme information. It has been found that some of the English multi-word phrases have no equivalent Telugu phrase available. Our system fails to identify some emotion words from their conjugated counterparts. For example, the Telugu word "ఒక (*oka*)" that means '*an*' is matching with "ఒకరికి కలిగిన మేలుకు కూడ సంతోషించు (*okariki kalingina meeluku kuda santhoshinchu*)" '*congratulate*' and "ఒక వృత్తిలోనివారు ఏర్పాటు చేసుకొన్న సంఘం" (*oka vruthilonivaaru erpatuchesukunna sangam*) '*Brotherhood*' which both belong to the *happy* list.

Affect Lists	**Hindi**		
Affect Lists	**Hindi**		
	Precision	**Recall**	**F-Score**
anger	87.65 **[*90.25*]**	76.64 [***80.53***]	81.22 [***85.50***]
disgust	81.88	58.89	70.23
	[***82.51***]	[**59.*77***]	[**71.*43***]
fear	85.22 [***90.11***]	60.86 [**70.58**]	67.49 [***80.21***]
joy	86.46 [***90.68***]	75.76 [**80.43**]	79.32 [***85.63***]
sadness	90.78 [***94.98***]	70.87 [***80.92***]	80.47 [***84.65***]
surprise	88.77 [***89.80***]	80.56 [**83.11**]	84.26 [***86.48***]
	Bengali		
anger	63.12 **[*65.22*]**	66.43 [***69.91***]	64.80 [***68.22***]
disgust	39.67 [***42.78***]	47.40 [**49.*37***]	42.55 [**45.*32***]
fear	62.49 [***74.91***]	70.87 [**80.00**]	65.10 [***77.49***]
joy	88.67 [***92.86***]	74.88 [**80.77**]	82.59 [***86.78***]
sadness	90.09 [***93.56***]	60.32 [***70.80***]	75.94 [***81.07***]
surprise	82.43 [***89.08***]	81.00 [**83.10**]	81.74 [***86.90***]
	Telugu		
anger	93.42 **[*95.45*]**	66.66 [***66.66***]	77.80 [***78.50***]
disgust	89.55 [***92.05***]	43.03 [4***9.39***]	58.13 [**64.*28***]
fear	92.43 [***94.49***]	40.00 [**60.21**]	55.83 [***73.55***]
joy	86.07 [***92.68***]	64.70 [**70.58**]	73.87 [***80.13***]
sadness	94.49 [***94.98***]	50.22 [***50.22***]	65.58 [***65.70***]
surprise	98.03 [***99.50***]	91.12 [**93.50**]	94.44 [***96.40***]

Table 3: Precision, Recall and F-Scores (in %) of the baseline systems for three target languages (Hindi, Bengali and Telugu) per emotion class on the translated *SemEval 2007* test corpora before and **[after]** including morphology.

Language	**Precision**	**Recall**	**F-Score**
Hindi	88.04	92.56	90.30
Bengali	75.83	71.66	73.33
Telugu	96.54	64.66	80.66
English	74.24	64.38	69.28
Japanese	83.52	49.58	62.22

Table 4: Comparative results of Average Precision, Recall and F-Scores (in %) of the three Indian languages (Hindi, Bengali and Telugu) with other two foreign languages (English and Japanese)

Language	Precision	Recall	F-Score
Emotion Score (ES) ≥ 0			
Hindi	88.04	92.56	90.30
Bengali	75.83	71.66	73.33
Telugu	96.54	64.66	80.66
Emotion Score (ES) ≥ 10			
Hindi	88.04	92.56	90.30
Bengali	75.83	71.66	73.33
Telugu	96.54	64.66	80.66
Emotion Score (ES) ≥ 30			
Hindi	88.04	92.56	90.30
Bengali	75.83	71.66	73.33
Telugu	96.54	64.66	80.66
Emotion Score (ES) ≥ 50			
Hindi	88.04	92.56	90.30
Bengali	75.83	71.66	73.33
Telugu	96.54	64.66	80.66

Table 5: Average Precision, Recall and F-scores (in %) for three target languages on different ranges of Emotion Scores

6. Conclusion

The present paper describes the preparation of *WordNet Affects* and basic prototype emotion analysis systems for three Indian languages (Hindi, Bengali and Telugu). The automatic approach of expansion and translation reduces the manual effort in building the lexicons and corpora. Though perfect sense mapping between languages is impossible, the comparative results with English and Japanese reveal that the emotion in text preserves its senses across languages. The resources are still being updated with more number of emotional words to increase the coverage. Our future task is to integrate more resources so that the number of emotion word entries in each of the target languages can be increased. The sense disambiguation task needs to be improved further in future by incorporating more number of translators and considering their agreement into account. The future task is to adopt language dependent and independent features for extending the emotion analysis task in multilingual platform.

7. Acknowledgements

The work reported in this paper was supported by a grant from the India-Japan Cooperative Programme (DSTJST) 2009 Research project entitled "Sentiment Analysis where AI meets Psychology" funded by Department of Science and Technology (DST), Government of India.

8. References

Ahmad, K., D. Cheng, and Y. Almas. (2006). Multilingual sentiment analysis of financial news streams. *In Proc. of the 1st Intl. Conf. on Grid in Finance*.

Aman, S. and Szpakowicz, S. (2007). Identifying Expressions of Emotion in Text. *V. Matoušek and P. Mautner (Eds.):TSD 2007, LNAI 4629*, 196–205.

Baccianella Stefano, Esuli Andrea and Sebas-tiani Fabrizio. (2010). SentiWordNet 3.0: An Enhanced Lexical Re-source for Sentiment Analysis and Opinion Mining. *In Proceedings of LREC-10, 7th Conference on Language Resources and Evaluation*, pp. 2200-2204.

Banea Carmen, Rada Mihalcea, Janyce Wiebe, and Samer Hassan. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of EMNLP'08*.

Banea Carmen, Rada Mihalcea, and Janyce Wiebe. (2010). Multilingual subjectivity: Are more languages better? In *Proceedings of COLING'10*.

Banea Carmen, Rada Mihalcea, and Janyce Wiebe. (2011). Sense-level Subjectivity in a Multilingual Setting. In *Proceedings of SAAIP Workshop, IJCNLP 2011, Chiang Mai, Thailand*, pp 44-50.

Bautin Mikhail, Lohit Vijayarenu, and Steven Skiena. (2008). International sentiment analysis for news and blogs. *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM-2008)*, Seattle, Washington.

Bobicev Victoria, Maxim Victoria, Prodan Tatiana, Burciu Natalia, Anghelus Victoria. (2010). Emotions in words: developing a multilingual WordNet-Affect. *CICLING 2010*.

Blitzer John, Mark Dredze, and Fernando Pereira. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *In Proceedings of the 45th Annual Meeting of the Association of Computational (ACL-2007)*, pp. 440–447, Prague, Czech Republic. Association for Computational Linguistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, vol. 20, pp. 37–46.

Das Dipankar and Sivaji Bandyopadhyay. (2009a). Word to Sentence Level Emotion Tagging for Bengali Blogs. *In Proceedings of ACL-IJCNLP-2009*, pp.149-152. Suntec, Singapore.

D.Das and S.Bandyopadhyay. (2009b). Sentence Level Emotion Tagging. *In the proceedings of the 2009 International Conference on Affective Computing & Intelligent Interaction (ACII-2009).* pp. 375-380, Amsterdam, Netherlands.

Das Dipankar and Sivaji Bandyopadhyay. (2010). Developing Bengali WordNet Affect for Analyzing Emotion. *In the proceedings of the 23rd International Conference on the Computer Processing of Oriental Languages*, pp. 35-40, California, USA.

Denecke, K. (2008). Using sentiwordnet for multilingual sentiment analysis. *In Proceedings of the International Conference on Data Engineering (ICDE 2008), Workshop on Data Engineering for Blogs, Social Media, and Web*, volume 2.

Ekman Paul. (1992). An argument for basic emotions, *Cognition and Emotion*, 6(3-4):169-200.

Esuli, Andrea. and Sebastiani, Fabrizio. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining, *LREC*.

Esuli Andrea, Fabrizio Sebastiani, and Ilaria C Urciuoli. (2008). Annotating Expressions of Opinion and Emotion in the Italian Content Annotation Bank. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC-2008)*, Marrakech, Morocco.

Jordan L. Boyd-Graber, Philip Resnik: Holistic Sentiment Analysis Across Languages: Multilingual Supervised Latent Dirichlet Allocation. *EMNLP (2010)*, pp. 45-55

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.

Ku Lun-wei, Yu-ting Liang, and Hsin-hsi Chen. (2006). Opinion Extraction, Summarization and Tracking in News and Blog Corpora. *In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, number 2001, Boston, Massachusetts.

Maks Isa and Piek Vossen. (2010). Annotation scheme and gold standard for Dutch subjective adjectives. *In Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, pp. 1327–1334, Valletta, Malta.

Miller, A. G. (1995). WordNet: a lexical database for English. In *Communications of the ACM*, vol. 38 (11), November, pp. 39-41.

Nakagawa Tetsuji, Kentaro Inui, and Sadao Kurohashi. (2010). Dependency tree-based sentiment classification using CRFs with hidden variables. In *Proceedings of NAACL/HLT '10*.

Prettenhofer Peter and Benno Stein. (2010). Cross-language text classification using structural correspondence learning. In *Proceedings of ACL'10*.

Quan, C. and Ren, F. (2009). Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis. *Empirical Method in Natural Language Processing and Association for Computational Linguistics*, pp. 1446-1454.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. (2007). Learning Multilingual Subjective Language via Cross-Lingual Projections. *In Proceedings of the 45th Annual Meeting of the Association of*

Computational Linguistics (ACL-2007), pages 976–983, Prague, Czech Republic.

Schulz Julia M., Christa Womser-Hacker, and Thomas Mandl. (2010). Multilingual corpus development for opinion mining. In *Proceedings of LREC'10*.

Seki Yohei, David Kirk Evans, Lun-Wei Ku, Le Sun, Hsin-His Chen, and Noriko Kando. (2008). Overview of multilingual opinion analysis task at NTCIR-7. In *Proceedings of the NTCIR-7 Workshop*.

Sood S. and Vasserman, L. (2009). ESSE: Exploring Mood on the Web. *3rd International AAAI Conference on Weblogs and Social Media (ICWSM) Data Challenge Workshop*.

Strapparava Carlo and Mihalcea Rada. (2007). SemEval-2007 Task 14: Affective Text. *45th Aunual Meeting of Association for Computational linguistics*.

Strapparava Carlo and Valitutti, A. (2004). Wordnet-affect: an affective extension of wordnet, *In 4th International Conference on Language Resources and Evaluation*, pp. 1083-1086.

Torii Yoshimitsu, Dipankar Das, Sivaji Bandyopadhyay and Manabu Okumura. (2011). Developing Japanese WordNet Affect for Analyzing Emotions. *In the proceedings of the WASSA 2.011, ACL-HLT 2011*, pp. 80-86, Portland, Oregon, USA.

Wiebe, J., Wilson, T. and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 1(2).

Wan Xiaojun. (2009). Co-training for cross-lingual sentiment classification. In *Proceedings of ACL/AFNLP'09*.

Collecting spontaneous emotional data for a social assistive robot

Clément Chastagnol[1,2], Laurence Devillers[1,3]
[1]Department of Human-Machine Interaction, LIMSI-CNRS, France
[2]University of Orsay PXI, France
[3]University of Sorbonne PIV, GEMASS-CNRS, France
E-mail: cchastag@limsi.fr, devil@limsi.fr

Abstract

The French ARMEN ANR-funded project aims at building an assistive robot for elderly and disabled people. This robot is controlled by a VCA (Virtual Conversational Agent), interacting with the subjects in a natural, spoken fashion. We focus in this paper on the gathering of emotional data in interaction with the VCA (or her voice only in the first gathering). 77 patients have participated in the data collection. The data will be used for building an emotion detection system. The specific difficulty in this project lies in the large variety of user voices (elderly, pathological) and affective behaviors of the patient. A questionnaire on the acceptability of the VCA and the quality of the interaction is also analysed and shows that the interaction with the VCA was deemed positive by the subjects.

Keywords: spontaneous emotional data, emotion detection, acoustic features, social assistive robot, elderly and pathological voices

1. Introduction

Machines are bound to become increasingly more people-oriented. With advances in the field of assistive robotics and Human-Machines interfaces, the development of "social assistive machines" has been predicted. The term, coined by Feil-Seifer and Mataric (2005), defines a machine designed with two purposes in mind: to physically help and support people with physical impairments and to provide social interaction to the user, generally in the frame of a specific task (rehabilitation, coaching, everyday-life assistance...). While the physical embodiment represented by the robot is necessary for objects manipulation, the social function may be carried out by a Virtual Conversational Agent (VCA) displayed on a screen, so the social assistive machine does not absolutely have to be a humanoid robot. A lot of efforts has been put into the development of assistive robots, especially for elderly people (Graf et al., 2002). The development of social robots is more recent but has seen applications in therapy for autistic children (Robins et al., 2005). Finally, the research on social interactions with VCA has been trying to tackle the problem of natural and multimodal interactions and of the evolution of engagement of the user across time in various applications as real-estate agent or exercise coach (Cassell, 2000; Bickmore et al., 2005).

We present in this paper the collection of an emotional corpus, using a VCA in interaction with the subjects. The VCA is to control a robotic platform. Section 2 describes the specifications of the whole system. In Section 3, details on the experimental protocol needed to acquire spontaneous emotional data are given. The collected corpora are presented in Section 4. Section 5 presents the results of the questionnaire on the acceptability of the VCA and the quality of the interaction and the conclusions are given in Section 6.

2. The ARMEN robot specifications

The French ARMEN ANR-funded project[1] aims at building an assistive robot for elderly and disabled people. It should be able to help people in everyday life by looking autonomously for lost and out-of-reach objects, handle them, and evolve in a realistic environment. Furthermore, it should call for help in case of emergency and behave as a companion: it should

[1] http://projet_armen.byethost4.com

understand smalltalk about specific topics and answer consistently with the emotion it detects. The interaction should be almost entirely conducted in a natural, spoken fashion with a VCA displayed on a screen. The communication system under development is composed of several modules: a speech recognition module, an emotion detection module and a dialog management module.

The development of these modules is a challenge because of the variety of user voices: some users have undergone surgical operations as tracheotomy and have difficulties producing a loud and clear voice. Most of the disabled people also have weak voices because they have lost control of their abdominal muscles. There is also the problem of noise generated by in-throat valves and the humming of respirators. Even the voices of elderly people in good health can be difficult to work with because it is sometimes unvoiced and whispered.

There are few publicly available corpora containing spontaneous emotional speech (Zeng et al., 2009). and even fewer with the types of voices found in this project. That is why it was decided to organize on-site data gatherings in medical facilities involved in the project. The choice of patients participating in this experiment was as wide as possible with regards to voice quality to see the most difficult cases. So far, two data gatherings have been conducted and more than 75 patients have participated in.

3. Experimental protocol and setup

Two data gatherings were organized in Montpellier, France in collaboration with the APPROCHE association[2] which promotes the use of new technologies for helping dependent people. The recordings took place in June 2010 and June 2011, adding up to eight days. Three medical facilities were involved: a functional reeducation center, a retirement home and a housing center for disabled people. The complementarity of these three sites allowed to record a broad spectrum of different, sometimes very marked voices.

The experiments included a Wizard-of-Oz system with an interviewer, a dialog module on a laptop and an operator triggering the dialog module unbeknown to the subject who thought he was having a real conversation with the module. The obtained reactions are thus close to a man-machine interaction in real context. For the first gathering, the subject was only interacting with a synthetic voice; a Virtual Conversational Agent was added for the second gathering.

The experiments were split into three phases: in the first

Figure 1: Screenshot of Mary, the VCA interacting with the subjects.

phase, the interviewer would present the project to the subject and explain the purpose of the experiment. The subject was invited by the interviewer to act emotions on purpose, by exaggerating the emotional tone of his voice.

In the second phase, the subject would interact with the dialog system in the frame of several scenarii (eight small for the first gathering, three longer ones for the second) designed to induce emotions by projection: the interviewer would explain the current scenario (a daily situation with an emotional potential) to the subject and ask him to imagine himself in the situation and to make the interface understand the emotion felt. The subject would interact with the dialog module operated by the accomplice. The accomplice made the dialog module answer according to pre-established strategies: to understand, to understand with empathy, not to understand, to be wrong. A dialog would set between the subject and the interface and would last on average 4 or 5 speaker turns per scenario for the first gathering and up to twenty for the second one.

The recordings were made using a wireless AKG lapel microphone with an M-Audio external soundcard. The sound was recorded in 32 bits, 16kHz mono WAV format with Audacity.

[2] http://www.approche-asso.com/

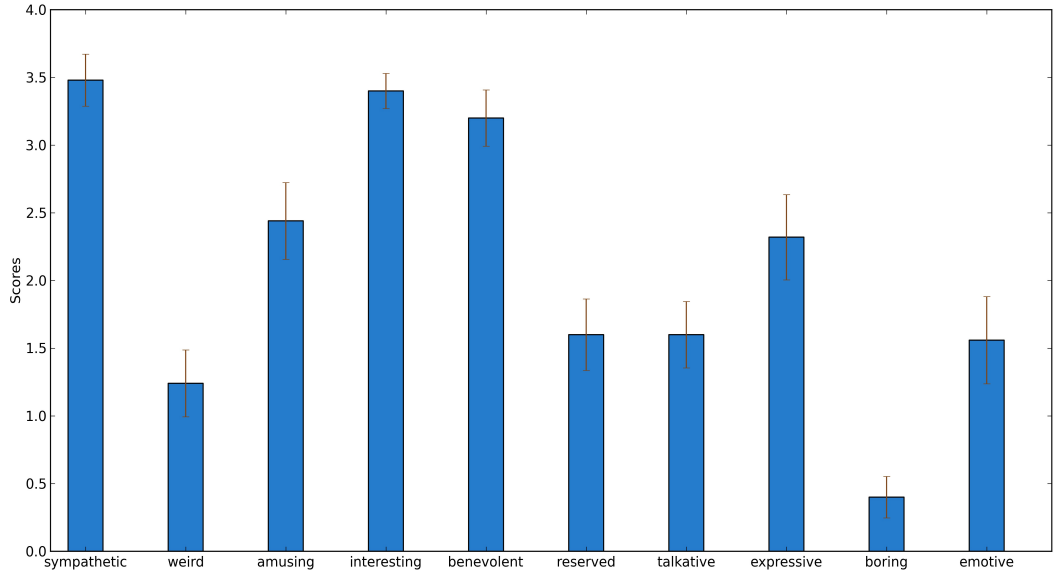

Figure 2: Attribution levels for several perceived traits of the VCA.

The scenarii were created jointly with members of the staff of the reeducation center and approved by physicians. They were inspired by daily-life situations in the centers and were meant to reflect the reality of what could be a user's experience of the robot. Both experiments were recorded and filmed; the sessions lasted between 9 and 37 minutes, with an average of 20 minutes.

In the third and final phase of the experiments, the subject would answer questions on the quality of the interaction, the acceptability of the VCA (or the synthetic voice) and their own personality.

Mary, the VCA used in the second gathering is based on the MARC platform currently developed at LIMSI-CNRS (Courgeon et al., 2008); it is pictured on Figure 1. It was controlled by a custom-made Wizard-of-Oz interface using the BML script language, which represents configurations of Action Units to animate the face of the VCA (Vilhjalmsson et al., 2007; Ekman & Friesen, 1978).

4. Presentation of the ARMEN corpora

The complete ARMEN_1 corpus for the first data gathering contains 17.3 hours of audio and video recordings from 52 persons from 16 to 91. The ARMEN_2 corpus for the second data gathering contains 8.7 hours of audio and video recordings from 25 persons from 25 to 91. The patients had no particular knowledge in computer sciences and experimental protocols.

The second phase of the experiment (scenarii) for both corpora was segmented and labelled by two expert annotators in emotionally-coherent segments using the following annotation scheme: 5 emotional labels (Anger, Fear, Happiness, Neutral, Sadness, plus an additional Junk label to deal with segments featuring distortion or microphone noises) and a scale of Activation on 5 levels.

Corpus name	ARMEN_1	ARMEN_2
# of segments	1996	1588
Kappa measure	0.33	0.37
% of segments kept	46%	63%
# of speakers	52	25
Class repartition:		
Anger	406 (20%)	92 (6%)
Fear	97 (5%)	21 (1%)
Happiness	427 (21%)	236 (15%)
Neutral	748 (38%)	1158 (73%)
Sadness	318 (16%)	81 (5%)

Table 1: Details on the ARMEN corpora.

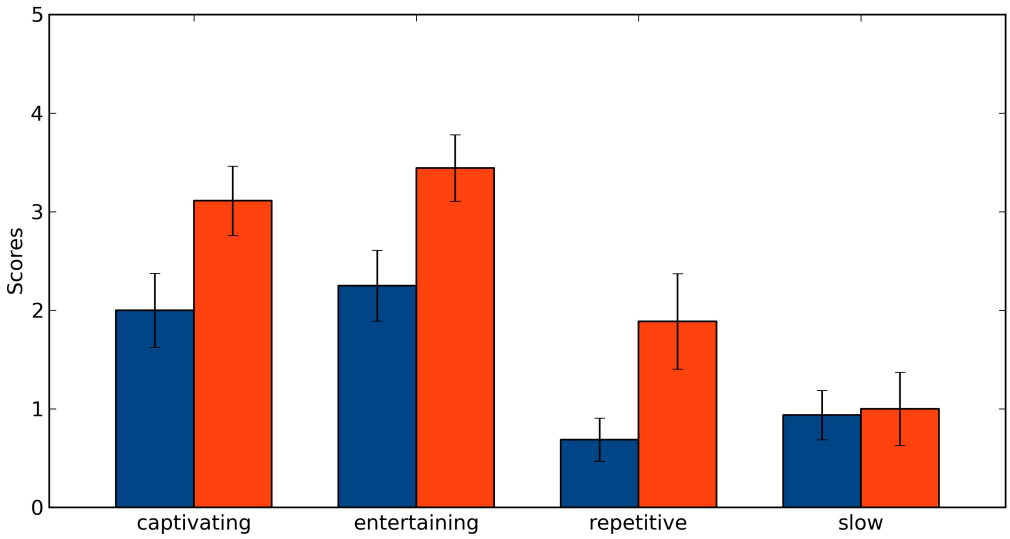

Figure 3: Results of the questionnaire for the quality of the interaction.

The segmentation followed a detailed protocol, yielding audio segments of at most five seconds, with a coherent emotional content. The two resulting corpora (ARMEN_1 and ARMEN_2) are detailed in Table 1 above.

5. Questionnaire results

Only the results of the questionnaire for the second experiment (with the VCA) are displayed in Figure 2 and 3 (standard error was used for the error bars). The questionnaire was designed to study two dimensions: the perception of the VCA by the users and the quality of the interaction.

The results show that the VCA was positively judged by the subjects of the experiment, with a high attribution of positive qualifiers and low attribution of negative qualifiers.

The quality of interaction was also deemed positive. There are however differences between the subjects of the second experiment, who came from a functional reeducation center and from a retirement home. The subjects coming from the former were younger and found the interaction captivating and entertaining, although a bit repetitive. They all declared to be willing to interact again with the VCA. The subjects from the latter were more reserved on this subject, but they found the interaction less repetitive and slow.

A surprising fact is that 95% of the subjects (even the older ones) preferred to interact casually rather than formally with the VCA, using their first name. A few even insisted on the fact that such a system could become as close as a family member some day.

6. Conclusion

We presented in this paper the collection of a corpus of spontaneous emotional data. More than 75 patients of medical facilities and retirement homes were put in interaction with a VCA to collect this data. We will use it to build the emotion detection module of a social assistive robot.

We consider this data very precious because of the variety of speaker voices and ages. This will allow us to try for mixes with other corpora of emotional voice and study differences in the detection of emotions according to age and voice quality.

Acknowledgement

This work is funded by the French ANR ARMEN project (http://projet_armen.byethost4.com). The authors wish to thank the association APPROCHE for their help during the data collection.

References

Bickmore, T., Caruso, L., Clough-Gorr, K., Heeren, T. (2005). It's just like you talk to a friend - Relational agents for older adults. Interacting with Computers,

17:711–35.

Cassell, J. (2000). More Than Just Another Pretty Face: Embodied Conversational Interface Agents. Communications of the ACM, Volume 43 Issue 4, pp 70 – 78.

Courgeon, M., Martin, J-C., Jacquemin, C. (2008). MARC: a Multimodal Affective and Reactive Character. In Proceedings of the 1st Workshop on Affective Interaction in Natural Environments.

Ekman, P., Friesen, W.V. (1978). The Facial Action Coding System: A Technique For The Measurement of Facial Movement. Consulting Psychologists Press Inc., San Francisco, CA, USA.

Feil-Seifer, D., Mataric, MJ. (2005). Defining socially assistive robotics. In Proc. IEEE International Conference on Rehabilitation Robotics (ICORR'05), Chicago, Il, USA, pp. 465–468.

Graf, B., Hans, M., Kubacki, J., Schraft, R. (2002). Robotic home assistant care-o-bot II. In Proceedings of the Joint EMBS/BMES Conference, Houston, TX, USA, volume 3, pp. 2343–2344.

Robins, B., Dautenhahn, K., Boekhorst, R., Billard, A. (2005). Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?. Universal Access in the Information Society (UAIS).

Vilhjalmsson, H., Cantelmo, N., Cassell, J., Chafai, N.E., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H. and van der Werf, R.J. (2007). The Behavior Markup Language: Recent Developments and Challenges. In Proc. of the 7th International Conference on Intelligent Virtual Agents, Springer.

Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S. (2009). A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 39-58.

A Ranking-based Emotion Annotation Scheme and Real-life Speech Database

Wenjing Han[⋆†], **Haifeng Li**[⋆], **Lin Ma**[⋆], **Xiaopeng Zhang**[⋆], **Björn Schuller**[†]

[⋆] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[†] Technische Universität München, Institute for Human-Machine Communication, Munich, Germany
wenjing.han@tum.de

Abstract

In this paper, we propose employing a learning-to-rank algorithm to the recognition of emotion in speech, and construct a novel ranking-based speech emotion recognition (SER) framework. We firstly design a ranking-based annotation scheme to collect high-reliability labels for model training. Next, we use the ranking scores to measure speakers' emotions and apply a learning-to-rank algorithm called ListNet to recognise (i. e., rank) emotion. A linear neural network is then trained for SER. Furthermore, a reference-based emotion visualisation approach is proposed to describe speakers' emotion fluctuation relative to a normal situation. Finally, feasibility of these methods is validated on the medium-scaled Mandarin real-life emotion corpus introduced for the first time which features massive 300 k individual pair wise comparisons.

1. Introduction

Speech is one of the most important information carriers in human-human communication. It conveys not only the linguistic information, but also the extra-linguistic information, including speaker identity, emotion, cultural background, etc. The emotional part of spontaneous speech plays an important role in the expression of speakers' stance towards the current topic in conversation and attitude to the conversational partners. In human-machine communication, such stance and attitude are extensively considered as guiding indicators for deciding on a machine's future course of action. Thus, it is essential for an intelligent human-machine interface to have the ability of speech emotion recognition (SER).

In the last few years, increasing attention has been paid to dimensional modelling in SER (Grimm et al., 2007; Giannakopoulos et al., 2009; Eyben et al., 2010a). This dimensional approach is an alternative to the more traditional categorical approach, which divides emotions into a fixed number of discrete categories. In contrast, the dimensional approach can implicitly offer an infinite number of emotion descriptions by representing emotions as points in a multi-dimensional emotion space. And, it is often confirmed to be a more suitable approach to recognise various emotions in real-life speech comparing to the categorical approach (Grimm et al., 2007; Eyben et al., 2010a; Gunes et al., 2011).

To recognise dimensional emotions, current works (Grimm et al., 2007; Giannakopoulos et al., 2009; Eyben et al., 2010a) adopt what we call the regression-based approach. This approach views the SER problem as a regression task, and employs regression models (e. g., Support Vector Regression (SVR)) to predict continuous-valued emotion primitives. In light of this, continuous-valued labels of speech utterances needs to be collected for model training. And for this collection, the rating-based annotation scheme is commonly used. Generally, the annotators are asked to directly rate their emotion perception of utterances either by selecting a point in a graphic emotion space, or by selecting a level from a given ordinal scale or moving a slider in real-time. However, such a scheme has two main drawbacks. Firstly, performing such rating is a heavy cognitive burden to the annotators. For instance, most of the time, the annotators even cannot provide a convincing reason for themselves when they place an utterance at an exact point in an emotion space. Secondly, the lack of a uniform rating scale decreases the comparability of rating scores between annotators. For instance, a score can deviate significantly for different annotators. Due to such heavy cognitive burden and bad comparability, the reliability of labels is inevitably reduced, resulting in severe performance degradation of emotion recognition.

To address the above issues, we devise a ranking-based annotation scheme. It is inspired by a phenomenon universally reflected by annotators in our previous test (Han et al., 2011). In contrast to directly assigning numerical values to an utterance, it is easier for annotators to qualitatively compare emotions by expressing the fact that one emotion is more positive or more exciting than the other. Benefiting from this new scheme, a rater's work is simplified to rank the utterances' emotions by making pairwise comparisons. After a certain emotion primitive's annotation, an ordered list of all utterances is obtained; meanwhile each utterance's ranking score can be calculated.

To validate the feasibility of our ranking-based annotation scheme, a medium-scaled real-life Chinese corpus was recorded, and annotated for Valence (i. e., how positive or negative the affect appraisal is) and Arousal (i. e., how high or low the physiological reaction is) dimensions by our team, which we believe to be valuable to the research community given its real-life nature and novel annotation scheme. In addition, only sparse resources are available to-date of emotional Chinese speech containing naturalistic affect display.

We model the SER problem as a learning-to-rank task and regard the ranking score as emotion degree indicator. Moreover, to satisfy the real time requirement of a human-machine interface, we choose a listwise approach to predict ranking scores for each utterance. Specifically, ListNet, an effective listwise learning-to-rank algorithm, is employed to train a linear neural network based ranking model for emotion recognition. Its good performance in describing relative relation between speech emotions has been validated in our experiment. In addition, we propose a concept of reference

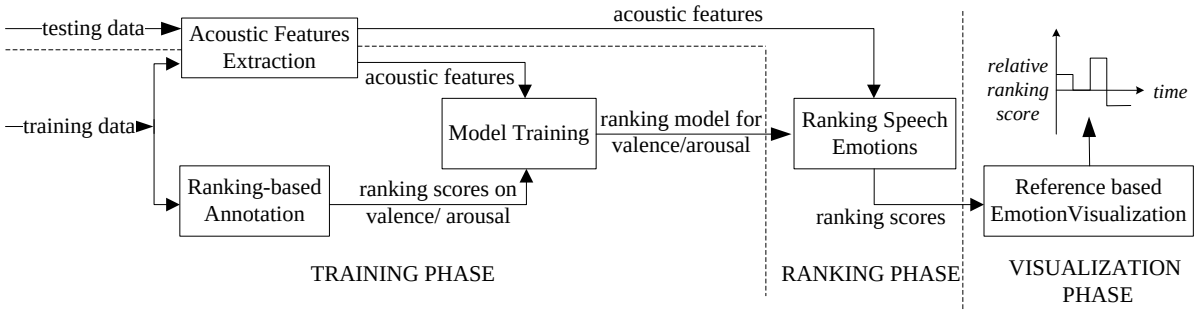

Figure 1: Schematic diagram of the ranking-based SER system.

emotion and visualise emotions in the form of a relative emotion fluctuation curve.

By that, we propose several methodologies to develop a ranking-based SER system covering the annotation, learning, and emotion visualisation processes. Despite the fact that a ranking approach has been used in the field of music mood classification (Yang and Chen, 2011) and affective analysis of movie scenes (Soleymani et al., 2008), it has – to our best knowledge – never been adapted for real-life SER before.

The rest of the paper is organised as follows: Section 2 describes our ranking-based annotation scheme. Section 3 details the ranking-based emotion recognition framework. Section 4 depicts the construction of our database and the extraction of acoustic features. Then, Section 5 presents our experimental method,results and discussions before concluding in Section 6.

2. Ranking-based annotation scheme

Speech emotion annotation is essential for constructing supervised SER systems. A good annotation scheme for speech emotions can provide the SER system with high-quality learning data, and by that enhance the accuracy of the system. Among criteria for a good annotation scheme two prevail: (i) describing speech emotions accurately, and (ii) being easy to implement.

In the field of SER, the most frequently adopted speech emotion annotation scheme in previous works is the categorical scheme. To design a categorical scheme, first one needs to select a fixed number of labels from the hundreds available, between which annotators are able to distinguish and reliably identify in utterances. However, there is little consensus regarding a standard taxonomy of emotional labels. Second, within a given taxonomy, the annotators are asked to assign the best-fit one to each utterance. Obviously, it is impossible to use a small number of labels to cover varied emotions in real-life speech. To describe real-life emotion more accurately, an alternative scheme must be employed.

A more fruitful alternative is the dimensional scheme, a technique in which emotions of utterances can be annotated as points in multi-dimensional emotion space (e. g., valence-arousal space). In this case, how to find the appropriate position for each utterance's emotion is a vital issue which needs to be considered. Most existing studies adopted rating measures to help annotators annotate emotions on a continuous scale. For instance, the scheme employed in (Grimm et al., 2007) asks annotators to select the best describing image for each emotional primitive (i. e., valence, activation, and dominance) after being played an utterance. The FEELTRACE instrument (Cowie et al., 2000) presents a colourful pointer and key emotion words to help annotators to select appropriate points in the valence-arousal plane and more advanced versions followed. However, there are also serious issues with the rating per se as discussed in Section 1, namely the cognitive burden issue as well as the low comparability issue, which tends to reduce the reliability of the annotations.

To address these issues, we design a ranking-based annotation scheme to help annotators comprehend a continuum of emotions in a comparative way. According to this scheme, pairwise comparisons of quadratic complexity $n(n-1)/2$ are required to obtain a straight ordering for n utterances. Specifically, in each round an annotator is presented with a pair of utterances which have not been compared with each other before by this annotator, and he or she is asked to select the more intense one afterwards – given the two dimension valence and arousal in the specific considerations that follow. Besides, the $equal$ option is available rendering the decision ternary. Finally, all of the pairwise comparison results need to be combined to produce one global ascending ordering per emotion primitive such as the named ones valence or arousal. Then, for each utterance, its ranking score is assigned to its ranking order (cf. Section 3.2).

It is a lengthy process to complete the $n(n-1)/2$ comparisons, especially in the case where n is large. Future efforts may thus focus on finding efficient ways of reducing complexity, e. g., by finding extrema and comparing to these as well as a given number of evenly spread samples in between maxima. At present, however, we spread annotation effort among several annotators in order to ease their cognitive burden of annotation. We argue that the reliability issue weighs more.

3. Ranking-based emotion recognition

A schematic diagram illustrating the training, ranking, and visualisation phase of our proposed SER system is shown in Fig. 1. In the training phase, the ranking-based annotation scheme described in Section 2 is performed to collect the labels for model training, and apply a listwise learning-to-rank algorithm: ListNet (Cao et al., 2007) as described in Section 3.2 is employed to train a ranking model; in the ranking phase, the ranking model predicts the ranking score for each input utterance; in the visualisation phase, the

relative emotion fluctuation curve is proposed to represent the recognition results (cf. Section 3.3).

3.1. Learning to rank

In this section, let us give a brief review on learning to rank and discuss the reason for choosing the ListNet algorithm to train the ranking model.

The task of learning to rank is to construct a model or a function for ranking objects. It has been successfully applied in document retrieval, collaborative filtering, and many other applications. There are two major approaches to learning to rank, respectively referred to as pairwise approach (Herbrich et al., 1999) and listwise approach (Cao et al., 2007).

The pairwise approach regards object pairs as instances in learning and formalises the problem of learning to rank as classification task. Specifically, it trains a classification model, which classifies pairs into two categories (correctly and incorrectly ranked), for ranking. A number of existing classification methodologies has been used to develop such a pairwise approach, such as Support Vector Machines (SVM) and Boosting whereby the classification models then lead to Ranking SVM (Herbrich et al., 1999) and RankBoost (Freund et al., 1998) .

The pairwise approach, however, is not suitable for our SER task, as the pairwise based ranking model cannot give the ranking score of an input utterance, unless all possible utterance pairs are compared. This way its application is limited in the human-machine interaction environment which has a critical real time requirement. In other words, we do model the SER task as a ranking problem; however, our ultimate purpose is not really to rank speech emotions, but to use the ranking scores to measure the emotional degree of speech.

Fortunately, the listwise approach successfully solves the above named time issue. Different from the pairwise approach, it takes the list of objects as instances in learning, and minimises the listwise loss between the reference list and the predicted one. Its ranking model assigns a ranking score to each object directly. Formally, it reduces the computational complexity from the pairwise approach's at $O(n^2)$ to $O(n)$. Therefore, in this paper, we utilise the listwise approach to predict ranking scores.

3.2. ListNet based emotion recognition

ListNet is a widely used listwise approach for learning to rank. It adopts a linear neural network as ranking model and uses gradient descent techniques to optimise a top one probability-based listwise loss function. In this section, we give a general description on ListNet based emotion recognition.

In training, a list of utterances $u = \{u_1, u_2, \cdots, u_n\}$ and its global ordering $r = \{r_1, r_2, \cdots, r_n\}$ are given, where n denotes the number of the training utterances, and $r_i\,(i = 1, \cdots, n)$ denotes the ranking of u_i. Then, u's corresponding reference ranking scores list $y = \{y_1, y_2, \cdots, y_n\}$ can be defined as follows:

$$y_i = r_i. \tag{1}$$

This definition ensures that the more positive or excited utterances can be assigned higher ranking scores. An acoustic feature vector $x_i = [x_{i1,}x_{i2}, \cdots, x_{im}]^T$ is extracted from u_i, where m denotes the number of features. The list of feature vectors $x = \{x_1, x_2, \cdots, x_n\}$ and the corresponding list of ranking scores y then form the learning instance.

A ranking model $f(\cdot)$ is needed to predict ranking scores for input utterances. Specifically, for each feature vector x_i, it outputs a ranking score $f(x_i)$. For the list of feature vectors x, we obtain a list of ranking scores $z = \{f(x_1), f(x_2), \cdots, f(x_n)\}$. According to the ListNet algorithm described in (Cao et al., 2007), the linear neural network is utilised as the ranking model,

$$f(x_i) = w^T x_i. \tag{2}$$

$f(\cdot)$ is trained by minimising the listwise loss function $L(y, z)$. To define the loss function $L(y, z)$, the top one probability $P(y_i)$ is utilised to transform a list of ranking scores into a probability distribution. It represents the probability of the utterance u_i being ranked on the top and is defined as follows:

$$P(y_i) = \frac{exp(y_i)}{\sum_{j=1}^{n} exp(y_j)}. \tag{3}$$

With Cross Entropy as metric to measure the difference between the reference list and the predicted one, the loss function becomes

$$L(y, z) = -\sum_{i=1}^{n} P(y_i) \log P\big(f(x_i)\big). \tag{4}$$

This loss function is differentiable, so it can be minimised using gradient descent techniques. The gradient of $L(y, z)$ with respect to parameter w can be calculated as follows:

$$\Delta w = \frac{\partial L(y, z)}{\partial w} = \sum_{i=1}^{n} \Big(P(y_i) - P\big(f(x_i)\big)\Big) x_i. \tag{5}$$

Next, it is used to update the linear neural network's weight w,

$$w = w - \eta \cdot \Delta w, \tag{6}$$

where η denotes the learning rate.

Subsequent to the above gradient descent procedure, a linear neural network based ranking model is generated.

3.3. Reference-based emotion visualisation

It is considerably attractive for a SER system to have the ability of intuitive, transparent, and informative emotion visualisation. In this section, we propose a reference-based emotion visualisation approach to provide a visual description of how a user's emotion changes with time relative to a reference emotion. This approach is motivated by a general perceptive experience: when we say someone is in a certain emotional state, actually we have done a comparison between his or her current emotion with a reference normal emotion in our mind subconsciously.

Specifically, we define 'reference emotion' as the emotion in a normal situation and then use the relative ranking score (RelS) to measure the current emotion offset degree. The current emotion's RelS is calculated as follows,

$$RelS = CurS - RefS, \tag{7}$$

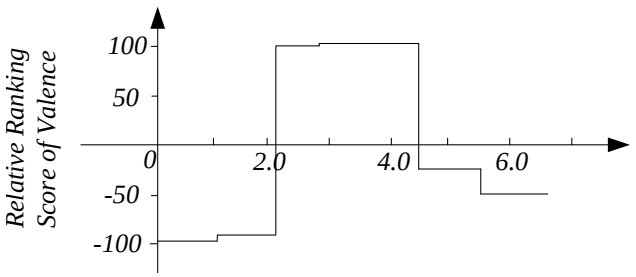

Figure 2: Valence REFC of an exemplary speech utterance sequence of the MREC database.

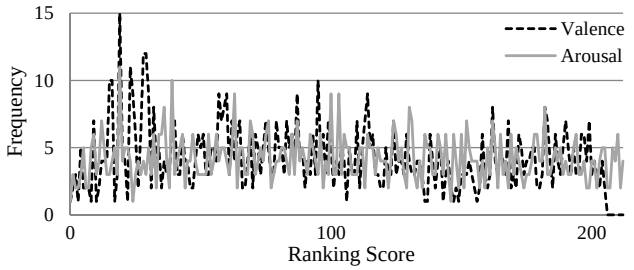

Figure 3: Emotion frequency at each ranking order in the MREC corpus.

where CurS and RefS denote the ranking score of the current emotion and reference emotion respectively. According to this approach, what we call the 'relative emotion fluctuation curve' (REFC) can be drawn. The valence REFC of a speech sequence of six utterances is shown in Fig. 2.

4. Database and Features Extraction

4.1. Data collection – The MREC Set

As there is no existing corpus featuring our proposed annotation scheme, we decided to go for annotation of a new corpus rather than re-label an existing language resource. In particular, there are two gaps present in the existing corpus landscape: (i) naturalistic emotions of spontaneous speech and (ii) under resourced languages – the lion's share is found for the Germanic language family by corpora as the corpora of the recent first challenges on Emotion at INTERSPEECH or AVEC – namely FAU AIBO Emotion Corpus, TUM AVIC, and AVEC (based on the SEMAINE corpus) or the popular Berlin and Danish Emotional Speech Databases and in particular for dimensional emotion the VAM corpus. Of the under resourced languages, a particularly interesting case are tonal languages given their use of intonation to encode linguistic meaning. We thus decided to record the 'Mandarin Real-life Emotion Corpus', or MREC for short. For the collection of utmost naturalistic every-day utterances, it was decided for a four week omni-present daytime recording set-up. Owing to ethical and privacy reasons the first author (female, 26 years, and native Mandarin speaker) of this paper volunteered as recording subject.

During a period of four consecutive weeks, she spent all her official working time in a quiet and isolated office space, where her daily spontaneous speech was recorded no matter what kind of communication channel she employed. Speech was recorded at 16 kHz, 16 bit quantisation using a close-talk microphone. The recordings are segmented and selected with a two-stage scheme manually: In stage one, speech with the same topic is kept in one segment, and only the segments with affective fluctuation as well as the subject's permission could be contained in our final corpus. At the end of this first stage, 39 topic-level segments were obtained. In stage two, each qualified topic-level segment was cut into isolated and context-sensitive utterances. At the end of this second stage, 890 utterances were collected. 528 of them stem from face to face conversations, and 362 utterances left stem from cell phone and voice over IP conversations. The average duration over all utterances is 1.4 seconds. While this corpus allows exclusively for speaker-dependent testing given a single speaker, we believe that there are many practical use-cases such as agent-human conversation (Eyben et al., 2010a) where a system has long-term interaction with a user. Moreover, here we are interested in a proof-of-concept on realistic data and ethical approval is a limiting factor throughout the collection of such private data.

4.2. Data Annotation

The manual annotation of speech emotion was accomplished by 12 annotators (8 male, 4 female) in terms of the ranking-based scheme as was described in detail in Section 2. Each annotator takes charge of about 32 967 pair wise comparisons for each dimension – i. e., of a twelfth of the total number of needed comparisons. Based on such an annotation, Fig. 3 shows the frequency of the emotions at each ranking order in the database. To ensure reliability of annotations, no limitation is given to the total duration of the annotation process and the annotators were allowed to listen to the utterances repeatedly.

4.2.1. Acoustic Features Extraction

For our experiment, the INTERSPEECH 2009 Emotion Challenge feature set (Schuller et al., 2009) of 384 acoustic features is extracted using our openSMILE toolkit (Eyben et al., 2010b).

5. Experiments and Results

5.1. Experimental method

The purpose of our experiment is to build a ranking-based SER system and validate the feasibility of the proposed methods. Specifically, the ListNet algorithm will be employed to train a ranking model for the recognition of emotion in speech. Moreover, by predicting real values, a regression model is also solving a ranking problem. Therefore we also adopt the k-Nearest Neighbor (k-NN) and Support Vector Regression (SVR) with radial basis function as kernel for performance comparison, as these regressors are commonly used for dimensional SER (Grimm et al., 2007; Eyben et al., 2010a). The systems adopt the same global ordered list as input. The same recognition experiments are carried out for valence and arousal respectively. For the sake of justified comparison, we ran a series of experiments to evaluate the performance by using ten-fold stratified cross-validation.

5.2. Results and discussions

In our experiment, (Pearson's) correlation coefficient (CC) and Gamma statistic G are adopted to measure the accuracy

Method	Learning Type	CC(V/A)	G(V/A)
k-NN	regression	.33/.76	.04/.03
SVR	regression	.53/.80	.04/.03
ListNet	learning-to-rank	.40/.78	.09/.06

Table 1: The Gamma Statistic of different learning algorithms for Valence (V) and Arousal (A) Recognition.

of speech emotion recognition. CC is a widely used measure in the SER field, and G is defined as follows,

$$G = \frac{C - D}{C + D}, \tag{8}$$

where C denotes the number of correctly ranked pairs, and D denotes the number of incorrectly ranked pairs. G equals 1 for perfect agreement, -1 for total disagreement, and 0, if the rankings are independent.

These two measures of the different named methods are shown in Table 1. It can be observed that, the CCs of arousal recognition are obviously higher than those of valence. This is a well-known typical behaviour as arousal is usually well-assessed by acoustic descriptors, whereas valence benefits from linguistic or additional facial expression information (Grimm et al., 2007) (Eyben et al., 2010a). Further, the perception of valence is usually more ambiguous than that of arousal. The interesting part is that, the Gs just show an opposite situation, which indicates more utterances are ranked correctly in valence than in arousal. This table also shows that even though ListNet does not perform to full satisfaction in CC, it achieves the highest value for G. This indicates that, ListNet performs better than the SVR methods in describing the relative relation between emotions. Note that, this gain is significant at a level of $p > 10^{-3}$ employing a one-sided z-test for both dimensions. This is reasonable, since the SVR methods are originally designed to make predicted values around the corresponding actual values as close as possible, but not to ensure the correct order between predicted values.

6. Conclusions

In this communication, we proposed a novel ranking-based SER framework. In contrast to the current state-of-the-art regression-based solutions, its contributions mainly manifest in four aspects: (i) to replace the current low-reliability rating-based annotation scheme, a ranking-based scheme was designed, which simplifies the annotation process and by that alleviates human annotators' cognitive burden. These in turn presumably enhance the reliability of annotation results and improve the performance of SER systems; (ii) we further introduced the MREC data-set of naturalistic emotion in the under-resourced mandarin tonal language featuring a fine-grained ranking based on massive 300 k individual comparisons open to the community per request; (iii) in the recognition model building phase, with the formalisation of the SER problem as a ranking task, a listwise ranking model was utilised to replace the current most frequently adopted regression models which were shown less suited to describe the relative relation between emotions in our experiments, and, furthermore, the ListNet algorithm was employed to build a linear neural network based ranking model; (iv) a reference-based emotion visualisation approach was proposed, which focuses on the description of users' emotion offset relative to normal situation.

As to future work, we plan to validate our approach on further corpora, and find more efficient ways to reduce the complexity of the annotation scheme, which seems crucial for its success.

7. References

Z. Cao, T. Qin, and T. Liu. 2007. Learning to rank: from pairwise approach to listwise approach. In *ICML*, pages 129–136, Corvallis, Oregon, USA.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, Sawey M., and M. Schröder. 2000. FEELTRACE: An instrument for recording perceived emotion in realtime. In *ISCA Workshop on Speech and Emotion: Developing a Conceptual Framework*, pages 19–24,, Belfast, UK.

F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie. 2010a. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 3:7–19.

F. Eyben, M. Wöllmer, and B. Schuller. 2010b. openSMILE - the munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, Florence, Italy.

Y. Freund, R. Iyer, and R. Schapire. 1998. An efficient boosting algorithm for combining preferences. In *ICML*, pages 170–178, Madison, Wisconsin, USA.

T. Giannakopoulos, A. Pikrakis, and S. Theodoridis. 2009. A dimensional approach to emotion recognition of speech from movies. In *ICASSP*, pages 65–68, Taipei, Taiwan. IEEE.

M. Grimm, K. Kroschel, and S. Narayanan. 2007. Support vector regression for automatic recognition of spontaneous emotions in speech. In *ICASSP*, volume IV, pages 1085–1088, Honolulu, Hawaii, USA. IEEE.

H. Gunes, B. Schuller, M. Pantic, and R. Cowie. 2011. Support vector regression for automatic recognition of spontaneous emotions in speech. In *EmoSPACE, held in conjunction with FG 2011*, pages 827–834, Santa Barbara, CA. IEEE.

W. Han, H. Li, and L. Ma. 2011. Estimating continuous-valued emotion of real-life speech. *Journal of Convergence Information Technology*, 6(6):308–316.

R. Herbrich, T. Graepel, and K. Obermayer. 1999. Support vector learning for ordinal regression. In *ICANN*, pages 97–102, Edinburgh, UK.

B. Schuller, S. Steidl, and A. Batliner. 2009. The interspeech 2009 emotion challenge. In *Interspeech*, pages 312–315, Brighton, UK.

M. Soleymani, G. Chanel, and J. Kierkels. 2008. Affective ranking of movie scenes using physiological signals and content analysis. In *ACM workshop on Multimedia semantics*, volume IV, pages 32–39, Vancouver, British Columbia, Canada.

Y. Yang and H. Chen. 2011. Ranking-based emotion recognition for music organization and retrieval. *IEEE Trans. Audio, Speech, and Language Processing*, 19(4):762–774.

A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpora

John Snel*, Alexey Tarasov†, Charlie Cullen*, Sarah Jane Delany†

*Digital Media Centre, Dublin Institute of Technology
Aungier St, Dublin 2, Ireland
john.snel@student.dit.ie, charlie.cullen@dmc.dit.ie

†School of Computing, Dublin Institute of Technology
Kevin St, Dublin 8, Ireland
aleksejs.tarasovs@student.dit.ie, sarahjane.delany@dit.ie

Abstract

This paper demonstrates the use of crowdsourcing to accumulate ratings from naïve listeners as a means to provide labels for a naturalistic emotional speech dataset. In order to do so, listening tasks are performed with a rating tool, which is delivered via the web. The rating requirements are based on the classical dimensions, activation and evaluation, presented to the participant as two discretised 5-point scales. Great emphasis is placed on the participant's overall understanding of the task, and on the ease-of-use of the tool so that labelling accuracy is reinforced. The accumulation process is ongoing with a goal to supply the research community with a publicly available speech corpus.

1. Intro

As part of building a naturalistic speech corpora, annotators are required to label and index emotional episodes associated with the acquired speech. In most cases, rather small numbers of "expert" labellers are asked to participate in listening tasks; the assignment of gathering large numbers of annotators is rarely a principal research objective. Moreover, most research does not indicate explicitly what expertise the annotators have. Expert listeners are usually researchers who are part of the wider field of emotional research.

Emotion is an important aspect of communication between *all* humans. The method used to accumulate ratings in this paper is through the use of crowdsourcing, which has been suggested by Tarasov et al. (2010). It diverges from others as we focus on large-scale listening groups not depicted as "expert" annotators—we suggest equal validity between an *expert* and a *non-expert* annotator's emotional judgement. That is to say, we aim to accumulate judgment ratings from a broader sample population that are not necessarily familiar with emotion theory.

The paper is structured as follows. Section 2 describes the related work in crowdsourcing and emotional speech labelling. Aims of our research are stated in section 3, and section 4 covers the methods used for creating the rating framework. The preliminary results are covered in section 5, and section 6 concludes the report.

2. Related work

In this section, a brief outline is given of related work in the area of accumulating labellers for corpora, and the labelling methods that have previously been used.

2.1. Crowdsourcing

Crowdsourcing is the use of tasks outsourced to a large group of non-expert individuals (Howe, 2008). Typically, a large number of tasks are distributed across a population of raters, and there from the results of several task solutions are combined. In the context of labelling corpora, each asset is presented to several raters and labelled separately by each individual. The final label for the asset is some combination of these labels; take majority voting for example (Brew et al., 2010). Crowdsourcing has recently been used for the task of getting labels for different corpora in numerous domains such as machine translation (Ambati et al., 2010), computer vision (Smyth et al., 1995; Sorokin and Forsyth, 2008), and sentiment analysis (Brew et al., 2010; Hsueh et al., 2009). Crowdsourcing is a fast way to accumulate labels; for instance, the work of Snow et al. (2008) received 151 ratings per hour, while Sorokin and Forsyth (2008) reported a speed of 300 ratings per hour. Nevertheless, with sufficient number of raters the quality of labels remains high and comparable to that of experts (Ambati et al., 2010; Snow et al., 2008; Sorokin and Forsyth, 2008).

Support for using crowdsourcing with regard to rating emotional speech is shown in the work of Cowie and Cornelius (2003). According to them, it can be argued that emotional expertise does not necessarily correlate with emotional experience, suggesting that the wider, non-expert population can provide labels that are equally valid to those of experts, who are primarily used to perform rating of emotional speech assets in state-of-the-art research.

2.2. Labelling naturalistic emotional speech

An early example of work that highlights the complexities in labelling naturalistic emotional speech is on the Leeds-Reading database (Roach et al., 1998). Emotional annotation came to four levels. The first level used freely chosen everyday emotion labels; the second specified the strength of the emotion, together with a sign to indicate valence; and, the third and fourth described emotional episodes based on the individual's appraisal of the event. Understandably, they specified that the number of categories associated with an in-depth qualitative coding strategy will amount to smaller occurrences in each category.

The development of the Belfast Naturalistic database

(Douglas-Cowie et al., 2000) followed from the Leeds-Reading experience. Their focus was to develop a quantitative description. They developed "trace" techniques to evaluate, quantitatively, emotion as it changes over time along underlying affect dimensions—positive to negative and active to passive. They argued that quantitative measurement using the Feeltrace tool better estimated real consensus compared to categorical labels, because of the inclusion of similarity—rather than only identical—measures. As somewhat unexpected, dimensional ratings showed less individual differences compared to categorical ratings, showing closer agreement on the evaluation dimension. For the rating task, however, they acquired three trained raters to use the tool; therefore, for this study, which excludes the need for comprehensive training inappropriate for crowdsourcing (large-scale, non-expert listening groups), the methods are adapted to meet the relevant requirements.

A comprehensive labelling schema for the JST/CREST Expressive Speech Corpus (Campbell, 2006) also included a version of the Feeltrace tool—and noted that labellers understood the meaning and validity of the two dimensions. Further, they proposed three levels for labelling: state of speaker, style of speaker, and physical aspects of the voice. This comprehensive schema is data-driven and appeared to be necessary when listening to speech in context and over long segments. For example, they familiarised themselves with the speakers mannerism when labelling someones speech over a five-year period. Such a comprehensive scheme, however, is not suitable for short segments of speech found in this particular study's speech dataset.

The study by Grimm, Kroschel and Narayanan (Grimm and Kroschel, 2008) used a three-dimensional model—valence, activation, and dominance. Interestingly, they discretised the continuous dimensional scales into 5 classes.

3. Aims

The focus of this paper, as part of an ongoing corpus building project, is to provide labels based on how naïve listeners judge conveyed emotional dimensions (i.e. effect-type orientation (Cowie and Cornelius, 2003)), for speech extracted from a previously constructed naturalistic, mood induced, emotional speech dataset (Cullen et al., 2008). Listeners are asked to rate on two scales that represent the activation-evaluation space.

Considering there is no absolute "ground truth" in emotion labels, and given that an individual's impression of emotion in speech is subjective in nature, it is suggested here that the use of crowdsourcing is a convenient method for determining more robust consensual ratings.

To collect ratings from large-scale listening groups, the listening tasks are performed through an online listening tool. The tool has its focus on user-centred design (UCD), developed and tested keeping in mind ease-of-use, ensure adequate understanding for each scale, and encourage participation by minimising the requirements of personal details. Moreover, the tool aims to be suited for repeated use to accumulate continual ratings from all participants.

4. Methods

This section describes the methods used to obtain the speech data, the framework chosen to label it, the available tool for the labellers, and the validation of tool design.

4.1. Data acquisition

The designated naturalistic emotional speech corpus for labelling is constructed based on Mood Inducing Procedures (MIPs) (Gerrards-Hesse et al., 1994). With inevitable restrictions in obtaining truly natural material while at the same time isolating the desired speech signal from unwanted noise, MIPs provide for a convenient trade-off. In this dataset, the inducing methods were performed on participants in a controlled environment with soundproof isolation booths. The build of the corpora (Cullen et al., 2008) investigated 3 different experiments incorporating the MIP 4 group (Success/Failure and Social Interaction MIP) and the MIP 3 group (Gift MIP). It considered several critical factors. Amongst these were: authenticity of emotional content, demand effects[1], ethical issues, and audio quality. The speech clips have been extracted from 8 different MIP sessions, and a total of 160 speech clips were chosen from 16 different speakers (7m/9f).

4.2. Labelling framework

To avoid the issues with subjective category labels, the labelling framework used in this paper is the dimensional approach as it appears to be more suited for cross-studies in a wider context (Eyben et al., 2003). Our method is comparable to the Feeltrace tool (Cowie et al., 2000), as mentioned above, mainly because of the number and type of dimensions used. We employ two-dimensions: activation and evaluation. Our method differentiates from the Feeltrace tool in two major ways.

First, our method is renouncing time-continuous evaluation, i.e. trace labelling (see also the work by Grimm and Kroschel (2005)), and instead provides annotation for utterances of discrete periods of time (termed as *quantised* labelling (Cowie et al., 2011)). The speech utterances rated are of short length (~5 seconds), and we are assuming that within the speech segment no changes in emotion occur, and are thus kept constant (Busso et al., 2008). For this study, prioritising large-scale rating via crowdsourcing is at odds with trace labelling that necessitates trained labellers.

Second, participants are presented with two discretised scales (colour-coded) rather than a continuous circular—or square—representation of the evaluation/activation space.

4.3. Design of web-based tool

To assist crowdsourcing, the rating tool is delivered via the Internet. The objective of the tool[2] is to have a simple but clean interface to make it easy for participants to understand and use. The participant's understanding about each rating scale is given considerable importance. The tool includes a detailed instructions page about how and what to annotate. As a more straightforward representation of the

[1] Demand effects are those possibilities of the subject guessing the purpose of the procedure and hence act the desired emotion.

[2] The online tool can be found at http://dmcx.dit.ie/emovere

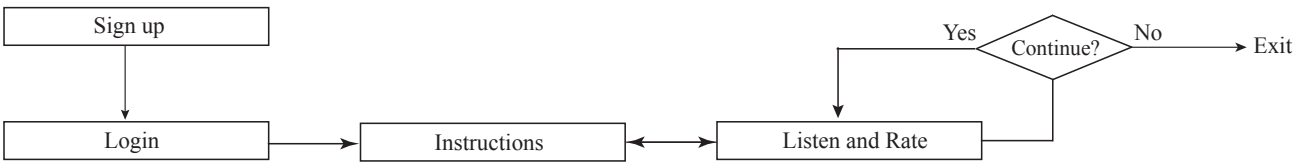

Figure 1: Flowchart of the presented web pages to the participant

circumplex model, whether circular or square, each dimension, activation and evaluation, is presented as two scales. For each scale, the participant is provided with a definition and an accompanied example. The design of the site (see Figure 1) ensures that the instructions are presented prior to the listening task, although the participants can refer back to the instructions at any stage during the task.

The participants are required to create a login account; and to prevent the impression of a daunting task and encourage participation, minimal details are required. However, mandatory information on first language and hearing impairment is required.

The listening task is presented as 3 successive steps i.e. listen to the speech clip and rate accordingly on both scales (see Figure 2). Each clip is only rated once by each participant. To avoid order effects, speech clips are randomised; and, to avoid fatigue and boredom effects participants are presented with just 6 speech clips before given the option to exit. Participants are given the option to skip a speech clip if they feel they cannot rate it by choosing "Do not rate". To prevent participants from continually doing this, it is required to fully listen to—or at least until the audio player has reached the end of the speech clip—before rating is activated. If a participant chooses "Do not rate" for 3 consecutive speech clips, they are notified and asked if they want to exit. A total of 160 speech clips are available for each participant to rate, and each clip can be replayed as many times as the participant wants. Participant details and rating information has been kept in two separate databases.

4.4. Preliminary survey (design validation)

Prior to implementation, we surveyed 7 non-expert (in emotional judgment) individuals to assess their understanding of the instructions using a multi-choice questionnaire. We ensured they were able to set up an account, and complete the task without difficulties. Participants were from a technical (college staff and other researchers) and non-technical (first year journalism students) background. The procedure for this was as follows:

1. Read instructions.
2. Answer questions about the definitions of both *evaluation* and *activation*.
3. Rate assets.
4. Assessment on workload.

For the activation question, 6 were correct and 1 incorrect; similarly, for the evaluation question, 6 were correct and 1

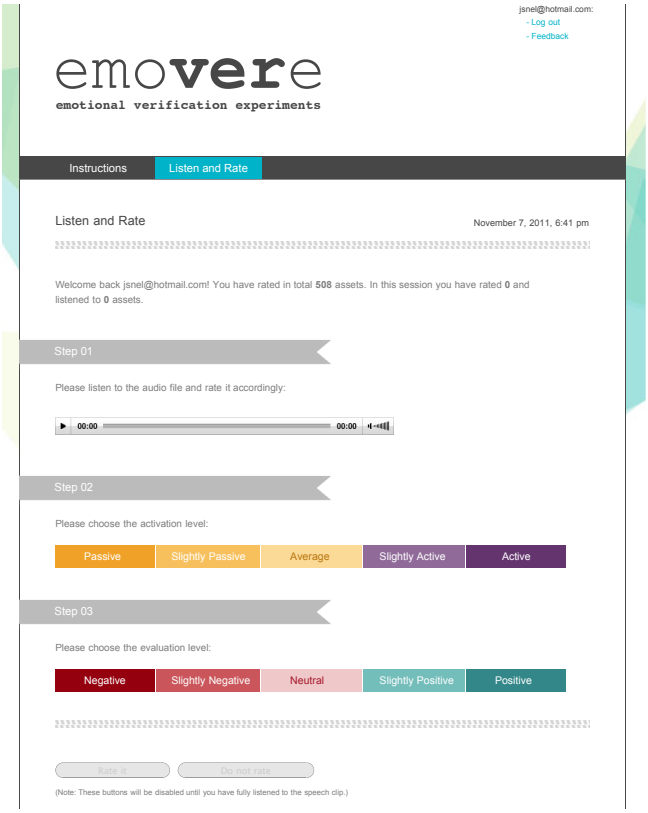

Figure 2: Online listening task

incorrect (see Table 1). It should be noted that the incorrect answers were from the same participant. The participant didn't follow the order of the above procedure. Instead, participant read instructions, rated assets, and then answered the questions on evaluation and activation. From this, it was concluded there was a sufficient amount of understanding among the raters for the instructions of both scales.

	Correct	Incorrect
Activation	6	1
Evaluation	6	1

Table 1: No. of correct and incorrect answers given for the multiple choice questions on activation and evaluation

A survey based on the NASA TLX (Hart and Staveland, 1988)—a subjective workload assessment tool—assessed the cognitive load on mental demands; temporal demands; and uncertainty, irritation, and stress (effort) while using

the online ratings tool. Overall, we concluded the cognitive demands were in adequate conditions (see Table 2).

Demand	VL	L	N	H	VH
Mental	1	1	3	2	0
Temporal	0	3	4	0	0
Effort	3	1	2	1	0

Table 2: Subjective workload assessment, VL=Very low, L=Low, N=Normal, H=High, VH=Very high

Participants were asked on the amount of assets that they would rate on a daily basis. 4 participants chose to keep it at 3 assets per day and 3 chose to increase the number. We concluded that participants should be presented with 3–7 assets at a time to prevent *boredom* and/or *fatigue* effects.
Besides querying cognitive load, participants gave free-response feedback on any other information they felt gave difficulties. Accordingly, technical issues within reason—such as browser issues and password restrictions—were addressed.
A brief summary of some interesting remarks from the free-respone feedback from the different participants is given as follows:

- Evaluation would be easier as binary.
- The definition of activation is easier to understand in terms of the dynamics of emotion.
- Scale for authenticity/genuineness could be introduced.
- There is a need for a baseline speech clip to compare others against.
- It was necessary to listen to some clips several times to hear the tone of voice, rather than the linguistic content.
- Others noted they assessed the clips along the scales according to the linguistic content.
- One participant said the speech clips were "weird".

5. Discussion

Since July 2011, we have received 1243 activation-evaluation pairs of ratings, which is 7.77 ratings per asset in average. The distribution of ratings for activation is shown in Figure 3, and for evaluation is shown in Figure 4. In total, 71 people have been registered as raters. Unfortunately, the majority have rated $<$20 assets, with a select few who provided labels for the whole corpus. The proportion of "Do not rate" ratings is only 3%, which shows that raters are rarely confused by the recordings. The evaluation dimension exhibits the same trend as would have been expected—it contains a large number of neutral ratings, gradually decreasing towards positive or negative classes. However, the corpus seems to have a relatively big number of active, non-neutral assets. One of the explanations can be the nature of the task faced by participants that forced them to act fast. In any case, it indicates that the MIP procedures used were successful in inducing non-neutral emotions.
We calculated the standard deviation (SD) for the ratings of each asset and used the mean value as a measure of rater agreement. The mean SD for the *activation* scale was 20% proportional to the width of the scale. Likewise, the *evaluation* scale came to 21%; that is to say, the participants are deviating from the average label by one class. The mean SD for this corpus was compared with the mean SD for the VAM corpus, which also used 5 discrete classes. The degree of agreement is comparable for both studies—VAM corpus is 14% for *activation* and 18% for *evaluation*.

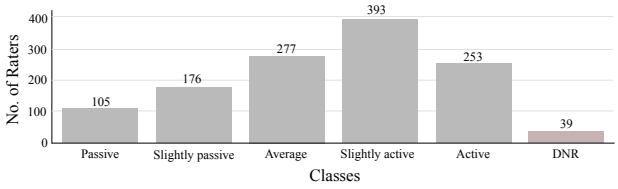

Figure 3: The number of ratings for the *activation* scale in the overall speech dataset, DNR=Do not rate.

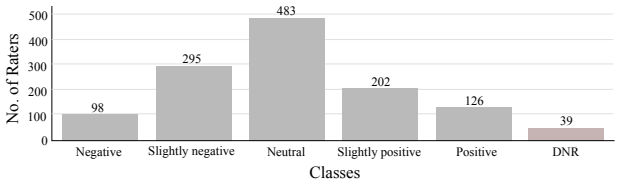

Figure 4: The number of ratings for the *evaluation* scale in the overall speech dataset, DNR=Do not rate.

6. Conclusions

One of our aims was to have participants engage with the tool on a daily basis, and rate six at a time to avoid fatigue and boredom effects that may cause spurious labelling. In spite of several reminders, it was difficult to achieve consistent daily rating from individual participants. However, the process of getting labels is on going. As an alternative, we are considering using crowdsourcing platforms such as Amazon Mechanical Turk[3] in addition to volunteer raters.
The release of all sets of ratings will be in the near future, including the single target label for each asset, obtained by aggregating the ratings submitted by raters. All rated assets will be freely available to the research community, with downloadable versions updated as ratings accumulate. With that, analysis on ratings will also be published. Finally, the corpus' speech dataset will be extended using other emotion eliciting methods, all in the same recording environment.

7. Acknowledgements

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253. Authors would like to

[3] http://www.mturk.com

thank Anna Deegan for the help with the implementation of the tool. We also express gratitudes to all raters, who participated in the research.

8. References

V. Ambati, S. Vogel, and J. Carbonell. 2010. Active Learning and Crowd-Sourcing for Machine Translation. In *Procs of LREC*, pages 2169–2174.

A. Brew, D. Greene, and P. Cunningham. 2010. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Procs of PAIS*, pages 1–11.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359.

N. Campbell. 2006. A language-resources approach to emotion: corpora for the analysis of expressive speech. In *The Workshop Programme Corpora for Research on Emotion and Affect Tuesday 23 rd May 2006*, page 1.

R. Cowie and R.R. Cornelius. 2003. Describing the Emotional States that Are Expressed in Speech. *Speech Communication*, 40(1-2):5–32.

R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schroder. 2000. 'FEELTRACE': An Instrument for Recording Perceived Emotion in Real Time. In *Procs of ITRW on Speech and Emotion*, pages 19–24.

R. Cowie, C. Cox, J. C. Martin, A. Batliner, D. K. J. Heylen, and K. Karpouzis. 2011. Issues in data labelling. In R. Cowie, C. Pelachaud, and P. Petta, editors, *Emotion-Oriented Systems. The Humaine Handbook*, Cognitive Technologies, pages 213–241.

C. Cullen, S. Kousidis, and J. McAuley. 2008. Emotional Speech Corpus Construction, Annotation and Distribution. In *Procs of LREC*.

E. Douglas-Cowie, R. Cowie, and M. Schroder. 2000. A new emotion database: considerations, sources and scope. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Citeseer.

Florian Eyben, A. Batliner, B. Schuller, D. Seppi, and S. Steidl. 2003. Cross-Corpus Classification of Realistic EmotionsSome Pilot Experiments. In *The Workshop Programme*, page 77.

A. Gerrards-Hesse, K. Spies, and F.W. Hesse. 1994. Experimental Inductions of Emotional States and Their Effectiveness: A Review. *British Journal of Psychology*, 85(1):55–78.

M. Grimm and K Kroschel. 2005. Evaluation of Natural Emotions Using Self Assessment Manikins. In *Procs of IEEE ASRU*, pages 381–385.

M. Grimm and K Kroschel. 2008. The Vera am Mittag German Audio-Visual Emotional Speech Database. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 865–868, Hannover, Germany.

S.G. Hart and L.E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload*, 1:139–183.

J. Howe. 2008. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Business.

P.Y. Hsueh, P. Melville, and V. Sindhwani. 2009. Data Quality from Crowdsourcing: a Study of Annotation Selection Criteria. In *Procs of ALNLP*, pages 27–35.

P. Roach, Richard Stibbard, Jane Osborne, Simon Arnfield, and Jane Setter. 1998. Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association*, 28(1-2):83–94.

P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi. 1995. Inferring Ground Truth from Subjective lLabelling of Venus Images. *Advances in Neural Information Processing Systems*, 7:1085–1092.

R. Snow, B. O'Connor, D. Jurafsky, and A.Y. Ng. 2008. Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Procs of EMNLP*, pages 254–263.

A. Sorokin and D. Forsyth. 2008. Utility Data Annotation with Amazon Mechanical Turk. In *Procs of IEEE CVPR*, pages 1–8.

A. Tarasov, S.J. Delany, and Charlie Cullen. 2010. Using Crowdsourcing for Labelling Emotional Speech Assets. In *Procs of W3C workshop on Emotion ML*.

Multimodal Analysis of Upper-Body Gestures, Facial Expressions and Speech

Serkan Özkul, Elif Bozkurt, Shahriar Asta, Yücel Yemez, Engin Erzin
MVGL, Koç University
Rumelifeneri Yolu, Sarıyer, Istanbul, Turkey
E-mail: serozkul@ku.edu.tr, ebozkurt@ku.edu.tr, sasta@ku.edu.tr, yyemez@ku.edu.tr, eerzin@ku.edu.tr

Abstract

We propose a multimodal framework for correlation analysis of upper body gestures, facial expressions and speech prosody patterns of a speaker in spontaneous and natural conversation. Spontaneous upper body, face and speech gestures exhibit a broad range of structural relationships and have not been previously analyzed together to the best of our knowledge. In this study we present a multimodal database of spontaneous conversation. Within this database to identify cross modal correlations, we first perform unsupervised temporal segmentation of each modality using hidden Markov model (HMM) structures. Then, we perform correlation analysis based on mutual information measure, which is experimentally computed over the joint histograms of recurrent temporal segments (patterns) of modalities.

Keywords: Multimodal correlation, mutual information, speech-gesture, multimodal database

1. Introduction

An ideal system for automatic analysis and recognition of human affective information should be multimodal, as the human sensory system is. The integration of multiple sources of information would enhance the power for achieving a reliable emotional recognition; hence building multimodal databases is considered a very important issue for affective computing research. While this need is clearly acknowledged within the research community, very few large multimodal databases are available. Most of the databases deal only with speech or facial expressions and even when considering few more complete multimodal databases available they mostly combine audio and visual (facial expression) information, very few add upper-body gesture information.

Moreover, naturalness of the emotional database is another issue. Acquiring realistic emotional data is a challenging task. In fact, many of the databases available ask the subjects to act or pose emotions in order to extract speech and facial related features. In the last years, this lack of naturalism has been severely criticized. Recent research is oriented towards inducing emotions of speakers (elicited databases) or collecting real-life data.

There have been a large number of studies on emotion and non-verbal communication of facial expressions and also on expressive body movements. Yet, these studies were mostly based on acted basic emotions. HUMAINE database is one of the most comprehensive multimodal databases, which was collected during the Third Summer School of the HUMAINE EU-IST project, held in Genova in September 2006. Acted emotional state recordings of anger, despair, interest, pleasure, sadness irritation, joy and pride incorporated facial expressions, body movements and gestures and speech. The database is also multi-lingual including recordings in languages English, French and Hebrew [1]. In another study, Gunes and Piccardi analyze upper body gestures during six acted emotional behaviors [2].

Some of the most successful efforts to collect new emotional databases have been based on broadcasted television programs. Some of these examples are the Belfast natural database [3], the VAM database [4] and the EmoTV1 database [5]. Likewise, movie excerpts with expressive content have also been proposed for emotional corpora, especially for extreme emotions (e.g., SAFE corpus [6]).

2. System Architecture of Multimodal Database Collection

Building the data corpora is crucial part of the applied scientific studies. The data corpus should provide the means for understanding all the aspects of a given process. It should also direct the development of the techniques toward an optimum solution by allowing for the necessary calibration and tuning of the methods and also give good means for evaluation and comparison. Having a well-designed data corpus (i.e. capturing both general and also particular aspects of a certain process) is of great help for the researchers in this field as it greatly influences the research results. In this paper we present in detail the process of building our multimodal data corpus in the Multimedia, Vision and Graphics Laboratory (MVGL) of Koç University.

2.1 Recording Equipment

At MVGL, an automated multi-camera motion capture system is available to collect and analyze multi-view video data, which is primarily used for human body motion modeling applications. The motion capture system that we developed is based on 3D tracking of the markers attached to the person's body in the scene by making use of the multi-stereo correspondence information from multiple cameras. Portable camera, Drift HD170, is used as head attached camera to record facial expressions. The Drift HD170 supports 1080p HD recording at 30 frames per second and weighs 138g.

2.2 Recording Settings

We use the four cameras of the MVGL 8-camera optical

motion capture system, which are placed on a truss structure on the ceiling, positioned three meters away to capture the upper body movements of the participant in all directions for 3D tracking. These cameras are placed like a curve of a semicircle to prevent any marker getting out of the field of vision. This setting also increases the multi-camera calibration accuracy.

Moreover, during the recordings participants wear a black motion capture suit that is covered with optical markers, which are clearly visible in all lighting conditions. Totally 8 markers are used in the upper body tracking process that are attached to 8 human upper body parts (head, chest, 2x upper arm, 2x elbow, 2x forearm). The participant stands in the middle of the room during the recordings. We give the participant some degree of freedom in moving, but this freedom is limited to moving in place.

Speech is recorded with a high quality Sony ECM-166BMP lapel microphone worn by the participant close to mouth, so that movement of the participant would not cause volume alterations in the recordings. Moreover, it is small enough that the microphone does not occlude body markers. Lapel microphone and upper body recording system are both connected to a high capacity server that manages the synchronization between audio and video information. Speech of the actor is recorded at 16K sampling rate and stored in wav formatted files.

In addition, a remote-controlled head-mounted HD camera captures participant's facial expressions. The camera is mounted on a helmet worn by the participant and it can capture whole facial gestures. Since mount system with cam is light-weighted, the participant can easily move and rotate his/her head in any direction comfortably. The head camera records audio as well, but it is only used for synchronization purposes with the upper body recording system.

Figure 1: Four camera views with head camera on the subject and a sample view from head mounted camera

2.3 Recording Environment

The recordings are performed in the MVGL multi-camera studio with good lighting conditions. A good lighting condition means that there is enough diffuse light to leave no shadows on the participants face. All cameras are focused on the participant and the head-mounted camera is just adjusted in order to get a clear view of face. Sample views from multi-camera and head-mounted system are given in Figure 1. The background and the floor of the room are covered with black color.

2.4 MVGL-MUB: Multimodal Upper-Body Database

The Multimodal Upper-Body (MVGL-MUB) Corpus that we have collected currently consists of 42 recordings from five pre-defined scenarios in Turkish with 7 participants from Turkey. A summary of the multimodal database is given in Table 1.

Scenario	Conversation (Monologue/Dialogue)	Head-Cam (Y/N)	Record Count
1	M	Y	16
2	M	Both	9
3	M	Both	7
4	D	N	6
5	D	N	4

Table 1: Summary of multimodal database, Average record length is 3-5 minutes per recording.

We have used different types of scenarios in our records to extract characteristic features from gestures. Each scenario is designed for natural and transparent interaction of participants within the recordings, so that gesture profile can be used in characterization and synthesis of gestures in synchrony with speech data. A summary of the scenario descriptions are given as following:

Actor Id	Gender (Male/Female)	Age	Background
1	M	20-25	M.Sc. Student
2	M	20-25	M.Sc. Student
3	F	20-25	M.Sc. Student
4	M	20-25	M.Sc. Student
5	M	20-25	M.Sc. Student
6	F	25-30	Ph.D. Student
7	M	40-45	Professor

Table 2: Attributes of the subjects acted in the scenarios

Scenario 1: Storytelling a memory
The first scenario consists of an exciting event that the subject/participant faced with. Each subject tells an incident about his/her experience avoiding pretended gestures in spontaneous manner.
Scenario 2: Storytelling a documentary
The second scenario is storytelling a documentary film. Each subject watches the same documentary film and talks about it to another person in front of the cameras.
Scenario 3: Storytelling a fairy tale
In the third scenario, subject is asked to read/watch a fairy tale from a text/video, and tells the story as he/she remembers.
Scenario 4: Conversation with an agent
In the fourth scenario, subject is given a text from phone call between an angry client and customer support service employee who is asked to act part of the client and forced the subject to make overpowered gestures.
Scenario 5: Watch & Tell
As for the fifth case, subject is expected to watch and tell thoughts about various previously prepared videos and images.

3. Multimodal Analysis

Our multimodal analysis system consists of feature extraction, unimodal segmentation and correlation analysis parts. First, we extract features for each modality. Then, we use HMM classifiers to segment patterns in an unsupervised fashion. Finally, we employ mutual information for the correlation analysis of these modalities.

3.1 Feature Extraction

3.1.1 Upper Torso

The motion capture process is a marker-based approach where a set of distinguishable color markers is attached to the joints of the participant. As a result of optical marker tracking process, 3D world coordinates for each frame and each marker are calculated by using the camera calibration parameters. After getting the 3D marker movement data, smoothing operations are applied to marker data to eliminate noises and it is converted to joint angles in Motion Builder.

For a given frame in the video sequence, a set of N images can be obtained from the N cameras. We define the upper body gesture feature vector, $\boldsymbol{f}_k$ for frame k to include the Euler angles associated with the 3D joint angles with their first difference:

$$\boldsymbol{f}_k = [\theta_k, \phi_k, \psi_k],$$

where θ_k, ϕ_k and ψ_k are the Euler angles in the x, y and z axes, respectively representing posture of the speaker at frame k. The resulting feature vector is 3 dimensional for each chosen joint of the speaker. We consider joint angles of 5 body parts (left arm, left forearm, right arm, right forearm and neck), eventually 15 dimensional feature vector $\boldsymbol{F}_k^u$ is used to represent upper body gesture for each frame:

$$\boldsymbol{F}_k^u = [\boldsymbol{f}_k^1, \dots, \boldsymbol{f}_k^5]^T,$$

where $\boldsymbol{f}_k^j$ is feature vector for j^{th} body joint in frame k. After extracting body gesture features, we have temporally segmented all records into short action chunks (3–6 seconds long) and clustered these chunks into $N_g = 6$ groups by using unsupervised clustering. Then we observed each labeled action and tried to predict generic attributes of the clusters. These action groups are described below;

Label:	Description:
a	Raise hand: raise any hand from low to high altitude
b	Instant hand moves: make sudden and major hand moves
c	Two hand contact: unites both hands on front
d	Stand by: actor does no action
e	Hands down: Moves hands from high to low altitude
f	Circular hand action: make circular hand motion

Table 3: Gesture labels and their explanation

Each action chunk is stored in our gesture dataset in the following format:

$$\boldsymbol{S}_i^g = \left(b_i^g, e_i^g, n_i\right),$$

where b_i^g and e_i^g are begin and end time of the gesture label n_i respectively.

We have observed and validated gesture labels by using our supervision tool Video Observer. This tool displays the action clips along with their labels so we monitored performance of the unsupervised clustering process. According to our observation, we repeated clustering process with better parameters which gives more meaningful clusters

3.1.2 Face

We use Active Appearance models (AAM) for facial feature tracking. AAM is composed of shape and texture components. In order to build an AAM, we first extract video frames demonstrating distinct facial expressions and then label these images with N points defining the main features. These N points describe the shape of a face on a frame, whereas texture component can be described as pattern of intensities or colors across an image patch. Given a set of labeled images, all shapes are aligned into a common co-ordinate frame. Then, principal component analysis (PCA) is applied. Similarly, to build the texture model PCA is applied to grey level information of shape-normalized images. Finally, shape and texture parameters are concatenated. Then, to eliminate any existing correlations again PCA is applied.

We use the AAM-API for the implementation and totally track 88 facial points, including eyes, eye-brows, nose, mouth and chin. Then, we calculate center points for each

eye by taking average of eye corner coordinates. The distance of eye-brow points to these centers are used as features. Additionally, we also calculate inner mouth point coordinates differences and concatenate to eye-brow feature set. Finally, we calculate the first order derivative of the feature vector and obtain 48 dimensional facial expression feature vector.

3.1.3 Speech Prosody

Voice characteristics at the prosodic level, including intonation, rhythm and intensity patterns, carry important clues for emotional states. In this study, pitch frequency, first derivative of pitch frequency is considered in speech labeling procedure. Pitch variations, during word pronunciation, indicate emphasis on the phoneme and based on these variations each pitch alteration is tagged automatically with the most suitable label m from $N_p = 6$ choices below:

Label:	Description:
H*	Peak Accent: a high pitch tone target on an accented syllable relatively to speaker's pitch range
L*	Low Accent: a low pitch tone target on an accented syllable relatively to speaker's pitch range
L+H*	Scooped Accent: a low tone followed by a relatively sharp rise on an accented syllable
!H*	Down Step High Tone: a clear step down onto an accented syllable from a high pitch tone
DEA	De-accented pitch
SIL	Silence: non speech gap with no intensity value

Table 3: Prosody labels and their explanation

The pitch alterations during each syllable or word pronunciations, is tagged with one of 6 labels above so a tag sequence with time intervals is obtained. We define speech prosody feature as $\boldsymbol{S_i^p}$ for i^{th} interval-tag pair.

$$\boldsymbol{S_i^p} = \left(b_i^p, e_i^p, m_i\right),$$

where b_i^p and e_i^p are begin and end time of the prosody label m_i respectively.

Speech tagging process is done both in manual and automated fashion thus supervised and unsupervised prosody datasets are formed. For supervised dataset, we manually listened and labeled speech by using Praat software which is a program for acoustic analysis. Automated speech tagging is managed by our program Automated Speech Togger which is based on AuToBI Java toolkit [7].

3.2 Correlation Analysis

HMM classifiers are used to define re-occurring patterns on the unimodal streams of upper body, face and speech prosody. Then we employ mutual information for the correlation analysis of the modalities. The unimodal patterns are considered as discrete random variables, and the mutual information between two modalities is computed over the discrete event counts.

3.2.1 Mutual Information

A possible procedure to correlate the structure of the upper body, facial expressions and prosody streams is to match the corresponding segmentations – for example by counting the co-occurrences. Since speech and body gestures or facial expressions may have temporal correlation, we also consider time lag between modalities.

In discrete case, mutual information to correlate sequence of independent realizations of random variables *A* and *B* is defined as:

$$I(A,B) = \sum_a \sum_b P(A,B) \log \frac{P(A,B)}{P(A)P(B)}, \qquad (1)$$

where *P*(*A*) and *P*(*B*) are marginal densities and *P*(*A*, *B*) is joint density. Mutual information is always non-negative, and zero if and only if the variables are statistically independent. The mutual information takes into account the whole dependence structure of the variables, not only the covariance.

In our case, random variable *A* represents our gesture class and *B* represents prosody class. Thus mutual information can be computed using a co-occurrence matrix of prosody and gesture modalities.

3.2.2 Bimodal Co-Occurrence

To analyze structural relation between modals, we have counted temporal co-occurrence of each prosody label with corresponding gesture label(s), considering a time lag τ. For each prosody entry S_i^p in the record, gesture labels are counted in the time interval $(b_i^p - \boldsymbol{\tau}, e_i^p + \boldsymbol{\tau})$. Hence, we formed ($N_p \times N_g$) co-occurrence matrices $\boldsymbol{M}$ for unsupervised/supervised gesture/prosody datasets.

To be able to compute the mutual information on a given co-occurrence matrix $\boldsymbol{M}$, we normalized the matrix with the sum of its entries. Thus each entry in the normalized matrix represents the joint probability *P(A, B)* of how both modalities occur together:

$$P(a,b) = \frac{M(a,b)}{\sum_{x \in B} \sum_{y \in A} M(x,y)}, \qquad (2)$$

Hence the normalized matrix represents the probability distribution function of bimodal co-occurrence. The sum of a prosody label row then provides us with the marginal density *P(B)* of the corresponding prosody label. In the same way, by summing a gesture label column, we obtain the marginal density *P(A)* of the corresponding gesture label. On this basis, mutual information values are computed on each dataset for different time lag configurations.

4. Experimental Evaluations

We have calculated mutual information for gesture and

prosody modalities as described in Section 3.2. To this effect, we have used the unsupervised/supervised prosody datasets along with the unsupervised gesture dataset. Besides, different time lag parameters have been employed in the experiments to identify the structural delay between body gesture and speech. We give the mutual information results in Figure 2.

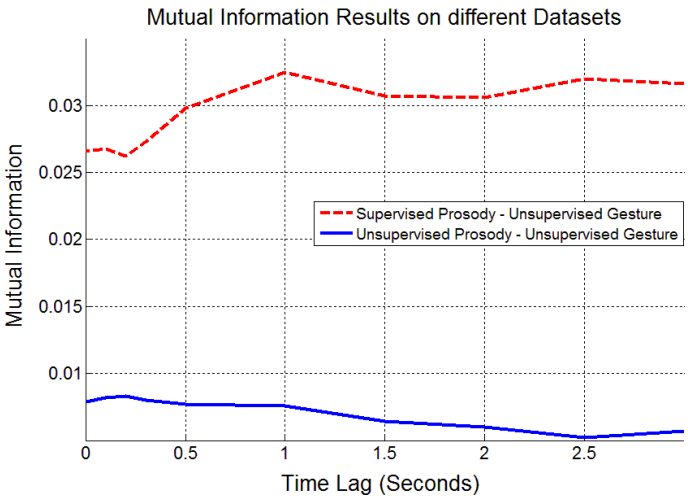

Figure 2: Mutual Information results for different datasets and time lag parameters

For our unsupervised prosody dataset, the highest mutual information value (0.008) is obtained with 0.3 second lag. We also observe that mutual information decreases as we increase time lag. In this dataset, prosody labels are automatically assigned based on pitch alteration. Also word structure or semantic is not considered in automatic speech tagging process. Thus wider and error-safe word intervals are tagged by our program. Since time-intervals for prosody labels are already long, increase in time lag parameter causes decrease in mutual information.

Alternatively, the supervised prosody dataset promises far better mutual information when compared to unsupervised version. The highest mutual information value (0.0325) is achieved in this case with 1 second time lag. We observe that a high mutual information value (0.0320) is obtained also at 2.5 seconds. Since prosody labels are assigned manually in this dataset, higher mutual information along with a more realistic lag value is obtained. Our experiments show that there is a lagged correlation between pitch alterations of speech and body gestures regardless of speech semantics.

5. Acknowledgements

This work was supported by Turk Telekom under Grant Number 11315-02.

6. References

[1] Humaine project portal: http://emotion-research.net/ (2011).

[2] Gunes, H., Piccardi, M.: Fusing face and body display for bi-modal emotion recognition: Single frame analysis and multi-frame post integration. In: Proc. of the 1st International Conference on Affective Computing and Intelligent Interaction, p. 102. Beijing, China, 2005.

[3] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases (2003).

[4] Grimm, M., Kroschel, K., Mower, E., Narayanan, S.: Primitives-based evaluation and estimation of emotions in speech. Speech Communication 49, pp. 787–800, 2007.

[5] Abrilian, S., Devillers, L., Buisine, S., Martin, J.: Emotv1: Annotation of real- life emotions for the specification of multimodal affective interfaces. In: Proc. 11th International Conference on Human-Computer Interaction (HCI 2005), pp. 195–200. Las Vegas, Nevada, USA, 2005.

[6] Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T.: The safe corpus: illustrating extreme emotions in dynamic situations. In: Proc. First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation), pp. 76–79. Genoa, Italy, 2006.

[7] Andrew Rosenberg, AuToBI project portal: http://eniac.cs.qc.cuny.edu/andrew/autobi (2012).

LAST MINUTE: a Novel Corpus to Support Emotion, Sentiment and Social Signal Processing

Dietmar Rösner(1), Jörg Frommer(2), Rico Andrich(1), Rafael Friesen(1), Matthias Haase(2), Manuela Kunze(1), Julia Lange(2), Mirko Otto(1)

1: Otto-von-Guericke Universität, Institut für Wissens- und Sprachverarbeitung,
Postfach 4120, D-39016 Magdeburg
{roesner, andrich, friesen, miotto}@ovgu.de
2: Otto-von-Guericke-Universität, Universitätsklinik für Psychosomatische Medizin und Psychotherapie
Leipziger Straße 44 D-39120 Magdeburg
{joerg.frommer, matthias.haase, julia.lange}@med.ovgu.de

Abstract

We present the LAST MINUTE corpus, a multimodal data collection taken from a carefully designed WoZ experiment that allows to investigate how users interact with a companion system in a mundane situation with the need for planning, re-planning and strategy change. The experiments have been performed with a cohort of N = 130 subjects, balanced in age, gender and educational level. The resulting corpus does not only comprise high quality recordings from audio, video and psychobiological signals, it contains as well transcripts from all interactions and data from a battery of well established psychometric questionnaires filled out by all subjects. For a subgroup of subjects audio records and transcripts from an additional post hoc interview are available as well.

Keywords: Corpus, Multimodal, Wizard-Of-Oz

1. Introduction

1.1. Goals and design rationale

"Really natural language processing" (Cowie and Schröder, Jan 2005), i.e. the possibility that human users speak to machines just as they would speak to another person, is a prerequisite for many future applications and devices. It is especially essential for so called companion systems (Wilks, 2010).

The interaction between a user and its personalized companion system will not always be without conflicts. A source for conflicts may e.g. be primarily located within the domain of discourse or it may be due to limitations of the technology resulting in some unnaturalness within the natural language user companion dialog.

Examples for the former type of conflict include situations where a companion has to insist to follow the higher goals of users: e.g. a diet companion that demands to obey dietary recommendations or a fitness companion that insists that the user performs his daily fitness program outside the house even when the weather conditions are not completely perfect if the user's higher goal is weight reduction or enhancement of physical fitness.

Conflicts may as well arise when e.g. during joint planning new information arises that partly or completely invalidates assumptions used so far resulting in the need for replanning or even strategy change.

The latter conflict may arise when a companion system is on the one hand a powerful conversant but falls on the other hand still short with e.g. more sophisticated inferences – that human dialog partners would draw without problems – or as long as the linguistic competence of the system is rich but still restricted.

The Wizard of Oz (WoZ) experiment that we report on here (Rösner et al., 2012) has been carefully designed with these considerations in mind.[1]

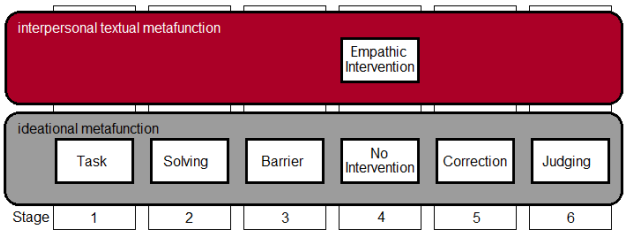

Figure 1: Stages of the WoZ experiment

1.2. Abstract view of the experiment

1.2.1. Systemic metafunctions

The perspective of systemic functional grammar (Halliday, 1976) is helpful in analysing conversation between users and companion systems. Halliday describes three metafunctions of language: the ideational, the interpersonal and the textual function. The ideational functionality is involved when content and facts (e.g. knowledge, perceptions and thoughts) are communicated. The interpersonal metafunction describes and creates roles and relations between the dialog partners. The textual metafunction relates utterances to the context and the situation.

During a conversation about some rational topic (e.g. in planning) negative emotions can occur – may be due to the problem or due to the dialog partner – that can not (or hardly) be handled only with the ideational metafunction. A companion system should be able to detect the emotional

[1] This is in contrast to our prior work with the NIMITEK corpus (Gnjatovic and Rösner, 2010) where affects and emotions were induced by WoZ simulating various malfunctions of a system.

state of its user or a problematic dialog situation and has to know a strategy to handle this situation. The companion system has to switch from the - in HCI normally dominant - ideational metafunction to the interpersonal and the textual metafunctions. Focussing these metafunctions the companion system should try to fix the emotional misalignment e.g. with an empathic intervention. When the emotional state of the user is detected as normal and stable again the companion system can switch back to the ideational metafunction to continue the dialog on the prior topic.

The LAST MINUTE experiment An compact view of the stages of the WoZ experiment (cf. fig. 1; for a more detailled description cf. (Rösner et al., 2012; Frommer et al., 2012)):

1. The companion system is first initialised by a personalization dialog. Then as the cover story of last minute the subject is told that he has won a holiday trip and may pack a suitcase for free. Some limitations such as the available time of the experiment are told, some are not. The cover story is provided with pictures and very detailed to stimulate ego involvement and imagination of the subject. The system's speech is focussed primarily on the ideational metafunction.

2. The subject starts with the actual task: choosing items from twelve categories while being supported by the system. The system alerts the subject when limitations such as time limit are infringed and provides information about the task's status - the remaining time.

3. Further information – about cold and rainy weather at the holiday destination – is provided that lets the subject realize that he had aimed a wrong goal and has now to change his strategy. This might lead to frustration and anger.[2] The subjects meet three barriers in this stage:

   (a) Listing: The suitcases content is listed and because of the length of this information the subjects may become impatient

   (b) Weight limit: In this stage the subject infringes the weight limit of the suitcase the first time

   (c) Weather: The information about the bad weather is given in this stage

4. The system addresses half of the subjects with an intervention using the interpersonal and textual metafunctions with increased intensity. This creates the chance for reflection and expression of anger. The intervention is designed to provide empathic help based on the principles of Rogers' paradigm of client centered psychotherapy (Rogers, 1959).

5. The subject gets the chance to revise a limited number of former decisions and repack the suitcase in this stage. The system's speech is focussed primarily on the ideational metafunction again.

6. The subject has to rate his own performance and if he is content with the content. One goal in this stage is to evaluate the style of attribution (self or external).

For additional descriptions see also (Rösner et al., 2011; Frommer et al., 2010)

Psychometric questionnaires After the experiment all subjects had to fill out a battery of paper-pencil-based psychometric questionnaires (for details cf. (Frommer et al., 2012)). The resp. data are part of the corpus and allow e.g. to correlate observed behavior and detected signs of affects and emotions with measured aspects of the personality of subjects.

2. The LAST MINUTE corpus

2.1. Naturalistic data

There is broad agreement that recording humans interacting in an environment of interest (e.g. SAL scenario (Douglas-Cowie et al., 2008) or companion scenario (Legát et al., 2008; Webb et al., 2010)) is a fundamental step towards assessing machine-human interactions within such scenarios (McKeown et al., 2010).

2.2. Other corpora

In table 1 we summarize various parameters of two widely employed corpora with naturalistic recordings – SAL (Douglas-Cowie et al., 2008) and SEMAINE (McKeown et al., 2010) – and contrast them with the resp. values for the LAST MINUTE corpus.
The SAL corpus contains recordings of WoZ simulated interactions with the four 'characters' of the so called Sensitive Artificial Listener (Douglas-Cowie et al., 2008), representing avatars with different combinations of personality features (e.g. Prudence, 'even-tempered and sensible' vs. Spike, 'angry and confrontational').
SAL may be seen as precursor of the SEMAINE corpus. The SEMAINE corpus has been used in the recent AVEC challenge (Schuller et al., 2011). As (McKeown et al., 2010) note 'the number of participants of the [...] SAL databases are too low to draw any general conclusions'. They therefore have extended this number. We have gone a step further by not only involving students but having a cohort that is balanced as well in age and educational level. As SAL and SEMAINE the LAST MINUTE corpus is available for the research community. The multimodal records in LAST MINUTE are available in high quality (cf. table 1 for details). The corpus is most useful for those interested in testing their developed affect or emotion classifiers on the contained naturalistic interactions or for those researchers that are interested in studying the more fundamental issues: what emotions and affects are expressed in which ways in realistic user companion interaction?

2.3. Transcripts

All experiments and interviews were transcribed by trained personnel following the GAT 2 minimal standard (Selting et al., 2009). These transcripts try to convey in written form as much information as possible from the audio records of the dialogs. All linguistic content of utterances is recorded

[2] We observed a broad variety of emotion expressions especially in this stage. See figure 2

Table 1: Comparison between corpora with naturalistic data: SAL (Douglas-Cowie et al., 2008), SEMAINE (McKeown et al., 2010) and LAST MINUTE (Rösner et al., 2012)

	SAL	SEMAINE	LAST MINUTE
Participants	4	20	130
Groups	students	students	balanced in age, gender, education
Duration	4:11:00	6:30:41	ca. 56:00:00 (+ ca 90:00:00 audio for interviews)
Sensors	2	9	13
Max. Video Bandwidth	352x288; 25Hz	580x780; 50Hz	1388x1038; 25Hz
Audio Bandwidth	20kHz	48kHz	44kHz
Transcripts	yes	yes	yes (GAT 2 minimal)
Biopsychological data	n.a.	n.a.	yes (heart beat, respiration, skin reductance)
Questionnaires	n.a.	n.a.	sociodemographic, psychometric
In depth Interviews	n.a.	n.a.	yes (70 subjects)
Language	English	English	German

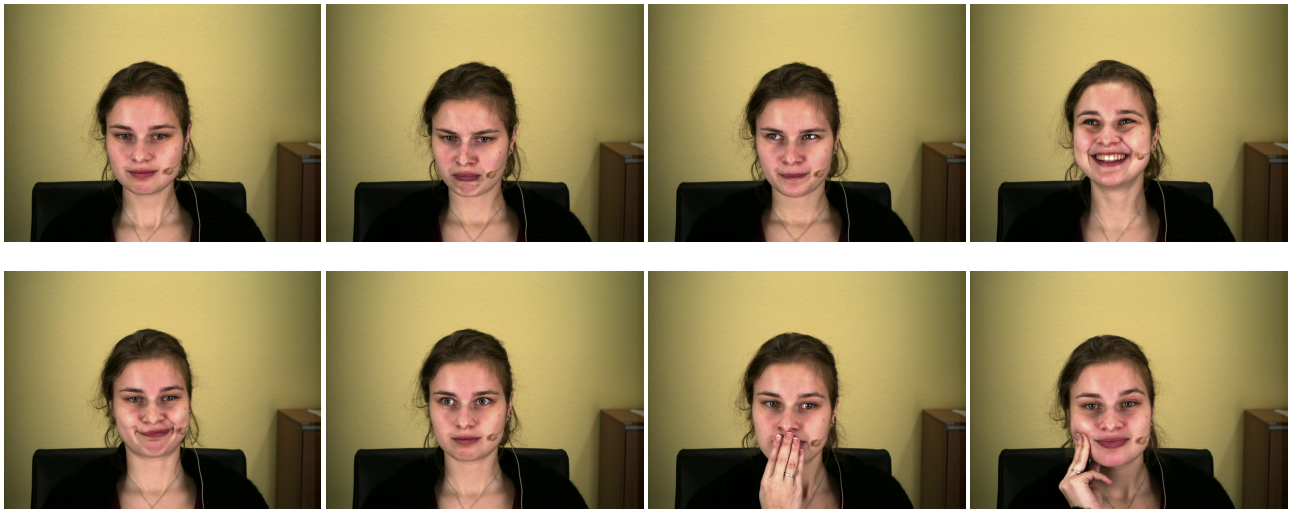

Figure 2: LAST MINUTE corpus: Different gestures and facial expressions of a single subject

in a manner as close as possible to the actual pronunciation of the speakers, so e.g. dialect is preserved. In addition acoustic social signals are transcribed in a descriptive (i.e. non-interpretative) manner. This includes all forms of interjections (e.g. 'hmmh', 'äh', ...) as well as pauses, breathing, sighing, laughter, etc.
All transcripts were made using FOLKER and EXMARaLDA (Schmidt and Schütte, 2010). Own software was employed to support the transcribers correcting misspellings, transcribing pauses and converting the wizard text.

3. Use cases, evaluation and results

The subjects were recruited with announcements, bulletins, and personal recruitment in e.g. choires. The experiments were conducted for nearly a year (August 2010 - June 2011). While conducting the experiments the transcription already started. The full set of transcripts for all N = 130 experiments is now completed and available for evaluation.[3]
Our own work in analyzing problem solving and linguistic behaviour of the subjects is based on the transcripts.[4] Other groups in our consortium SFB TRR 62 (Sfb, 2009) work on detecting and analysing emotional and affective cues in the recorded multimodal data from the LAST MINUTE corpus. This comprises analyses of:

- video signals: mimic and gestures (e.g. (Niese et al., 2010)) [5]
- audio signals: prosody, nonlinguistic events (e.g. (Scherer et al., 2010))
- psychophysiological signals: skin conductivity, heart rate, ...

In databases with acted emotion expressions often the classical emotions of Ekman (Cohn et al., 2002) are dominant and available in high intensity. In naturalistic records – like in the LAST MINUTE corpus – a completely different set of emotions and affective states is relevant. In addition the intensity is typically lower and there may be long phases without expressed emotions. This provides a challenge for

[3]For the audio files of N = 73 interviews the transcription will take some more time.

[4]Please note that this is the only part of the material that is dependent on the language of the subjects (i.e. German).

[5]Figure 2 shows indicative examples of visual expressions that are contained in the material and that are further investigated in this cooperation.

data analysis. A simple transfer of methods developed for acted material is not adequate and sufficient.
What emotions and affects are most relevant in user companion interaction (UCI) is part of the questions that are intensively investigated with the LAST MINUTE corpus. A still incomplete list of candidates includes e.g. surprise, boredom, irritation, frustration, helplessness, pride, shame, superiority, astonishment, feeling under time pressure, ...

3.1. Learning issues

The WoZ experiment is not primarily a learning experiment. Nevertheless the subjects have to solve a number of implicit learning tasks.
These tasks can be differentiated as follows:

- On the one hand subjects have to learn what the constraints are during their problem solving and with which actions they can be successful within the given constraints,
- on the other hand subjects have to learn how (i.e. with which lexical items and with which linguistic structures) they can effectively communicate with the system using speech in order to get their selected actions performed.

This is of course a more conceptual distinction. In the course of the interaction each single utterance of the subject typically may serve both purposes simultaneously.
In addition these learning tasks are embedded into the global choreography of the experiment. Learning is especially relevant, when barriers or challenges occur in the normal course of events.
Within the last minute module[6] the following learning tasks do occur with respect to problem solving and/or related to barriers and challenges:

- In the initial phase of the module the subjects get informed about the total number of categories and that in sum fifteen minutes are available for the packing of the suitcase. But the subjects are explicitly confronted with the local limit of one minute for each category only when they reach this limit in a category for the first time. As a consequence subjects might try to speed up their selection processes in subsequent categories.
- When the subjects get the contents of the suitcase listed verbally for the first time (after category six) they experience how time consuming such a listing is and if so - as a consequence - this increases their stress level. This may later in unpacking situations block them from asking for such a listing when they have to decide what items to unpack.
- Subjects sometimes try to stop such lengthy system utterances but the system does not allow barge-in.
- Subjects reach the weight limit barrier for the first time in category eight. Depending on their suitcase contents so far and their subsequent unpacking and packing they may reach this limit a number of times.
- The biggest challenge is reached in category ten when finally ('because of an interrupted data line') the information about the weather conditions at the target location Waiuku are available. In most cases this creates the need for a strategy change and a major repacking under increasing stress due to time constraints.

With respect to linguistic and dialog behavior during the last minute module subjects are implicitly confronted with questions like the following for which they have to find answers either through 'good guesses' based on their intuitions or through experimenting and trial and error:

- What linguistic structures and what lexical items can be used for packing commands, which words or forms are not available?
- Which techniques from human-human dialogs (e.g. anaphora, ellipses, ...) can be employed?
- How (i.e. with which lexical items and which linguistic structures) can a change of category be stipulated by the subject?
- What linguistic structures and what lexical items can be used for unpacking commands, which structures and items are not applicable?
- In a problem situation is there a way to get help? If so, how?

An example of a linguistic structure that subjects often try in packing that is not supported by the system is the use of conjunctive packing commands. When a subject utters a conjunctive command (e.g. 'I want three t-shirts and two jeans.') only the first conjunct is processed and acknowledged (e.g. 'Three t-shirts have been added.'). Subjects typically learn this constraint (One item type - with arbitrary number of instances - per command only) quickly and then stick to the constraint in subsequent utterances.

3.2. Effectiveness and efficiency

3.2.1. Wizard logs

The wizards have been trained and their behaviour has been anticipated and prescribed as nonambiguous as possible in a manual (Frommer et al., 2010).
All dialog contributions from the system (i.e. wizard) were pronounced by a TTS. The input for the TTS either was generated dynamically from the knowledge base (e.g. verbalisations of the current contents of the suitcase) or was chosen by the wizards from menus with prepared stock phrases. As a last option for unforeseen situations wizards could – supported by autocompletion – type in text to be uttered by the TTS.
In the course of more than 130 experiments with on average approx. 90 dialog turns each (in sum a total of ca. 11800 turns) only in one single turn – during the very first experiments – the wizards had to resort to the option of typing in an adequate wizard utterance because the prepared stock phrases did not suffice.
After a WoZ session all wizard contributions together with their timings are available as additional log file.

[6]The personalisation module of the experiment is independent from the last minute module with the domain of packing a suitcase

Evaluation of the wizard log files already allows to classify the overall interaction of different subjects with respect to a number of aspects. For other classifications NLP analysis of the contents of the subjects' utterances is necessary. (cf 3.3.)
An example of such an evaluation: In fig. 3 the result of a contrastive analysis of the dialog courses after the Waiuku barrier of all N = 130 subjects, divided into the subcohorts of elderly vs. young subjects, is given. The chart illustrates that more than half of the elderly have more negative dialog turns than the young subjects.
The differences between the groups are significant: a t-test yields a t-value of -3.78 and a p-value of 0.00024.

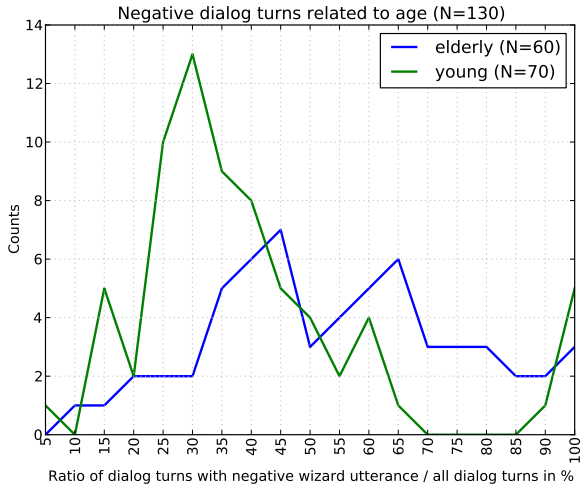

Figure 3: Contrastive evaluation of dialog courses after Waiuku barrier

3.2.2. Success and failure of dialog turns

Dialog turns of the subject are either successful or they may fail for a variety of reasons. In the following we will discuss failures during the last minute phase. Failed turns can easily be detected in the so called wizard logs because in the case of failure no confirmation is given by the system but some different utterance. Some system responses include hints about the cause of failure, others are unspecific and completely leave the problem of diagnosing the error to the user.

Success A dialog turn starting with a user request to pack or to unpack some items is successful when the situation allows to perform the requested action. In the wizard log file this can easily detected by the respective confirmative response of the system ('... wurd.* hinzugef.* ...', '... wurd.* entfernt ...').

Error messages The least specific 'error message' of the system tells the user that his utterance can not be processed ('ihre aussage kann nicht verarbeitet werden'). There are a number of reasons for using this 'catch all' system response. These include:

- The wizards conjure that the voice quality of the user's utterance is too poor for current ASR technology.
- The content of the user's utterance is beyond the allowed scope of the current subdialog.
- The syntactic or semantic complexity of the user's utterance is judged to be beyond the limits of current NLP technology.

The following system reactions are more specific:

- When a user tries to unpack items that have not been packed into the suitcase he gets the response that these items are not contained in the suitcase ('... nicht im Koffer enthalten'). Note that the system (i.e. wizard) demands for the exact terms from the menus as used during selection and does not accept synonymous terms. This is quite often a problem for subjects.
- When a user reaches the weight limit for the suitcase again then a packing command is responded to by the system with the message that the chosen item(s) can not be packed due to the weight limit (' ...k.*nn.* nicht hinzugef.* werden ...').
- When the time for a category is over (local time limit) then the system tells this to the user and enforces a change of the category (' ... muss jetzt beendet werden ...').

When users want to end the selection of items completely before the globally available time is over (e.g. by not choosing (further) categories for re-selection) then the system ask them the confirmation question if they are sure to end the selection already now ('... Auswahl an dieser Stelle beenden möchten ...').

A global measure In order to compare different dialogs we start with the following coarse global measure for the course of interaction of the last minute problem solving dialog: We distinguish turns that are - based on the logged system response – judged as successful from those that are judged as unsuccessful or faulty. We then use the ratio of unsuccesful turns in relation to all turns as measure of the relative faultiness of the dialog as a whole.
In detail:

- successful are all turns with an explicit confirmation of success,
- the turns that are counted as failed or unsuccessful include those with the following system responses (cf. above): unprocessable input, item not in suitcase, weight limit reached again, system enforced category change.

Data For a cohort of N = 130 subjects the values for this global measure range between 9 % and 73 % with a mean of approximately 26 % and variance 10.

3.3. Linguistic analyses

3.3.1. Evaluating transcripts

For the on-going evaluation of the records and transcripts of the interaction there are many interesting questions. In the following we will name just a few.

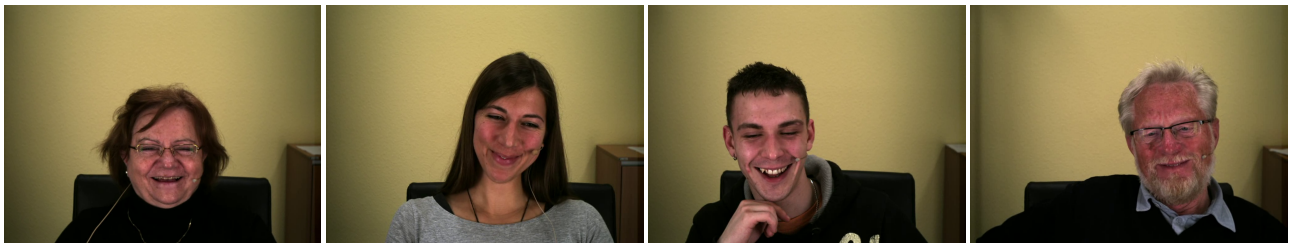

Figure 4: Examples of still images from the LAST MINUTE corpus: variations of laughing and smiling

The subjects in our WoZ experiments only get told that they will interact in spoken language with a new type of system that shall be personalised and that therefore will ask some questions and pose some tasks. They do not get explicit instructions about the linguistic constructions and interaction patterns that are possible or not possible to use in the interaction.

How do people with differing technical background interact in such a situation?

One might speculate that there are a number of different possible reactions in such a situation. People could on the one hand actively try to find out the possibilities and limitations of the system by starting with – from their perspective – simple constructions and then gradually further trying more complicated ones. On the other hand people could simply rely on their intuitive assumptions (their internal picture based on whatever prior experiences or projections) about such possibilities and limitations. In the latter case no active exploration will be tried and behavior – even if unnatural or inefficient – will be pursued as long as it proves to be successful.

Do users differ in their curiosity and there openness for experimentation? If so, how does this relate to results from the psychometric questionnaires?

Do users – although system initiative is dominant – at least try to take initiative by e.g. trying to ask questions? If so how do they react when the system is ignoring such attempts?

Do users e.g. try to interrupt lengthy system outputs (barge-in)?

Do users mirror the language of the system, e.g. on the lexical or syntax level? The system e.g. uses the general and somewhat uncolloquial term 'hinzufügen' (engl. to add) in its feedback for selection operations. Do users mirror this usage?

How quickly do users adapt to observed limitations of the system? An example: are there any re-trials once a user experiences that from a conjunction only the first conjunct gets processed by the system?

How quickly do users employ efficient communication strategies like the use of ellipses in consecutive dialogue turns? Are there any subjects that stick to full blown sentences in a sequence of dialogue turns with a high potential for ellipses?

Do users employ synonyms (or hypernyms or other semantic equivalents) when they choose objects from the selection menus or – even more to be expected – when they unpack items from the suitcase? How do they react when the system refuses to understand a semantically correct unpacking command that does not use the term from the selection menu but a semantically equivalent one (e.g. 'remove three pairs of socks' instead of 'remove three pairs of stockings')?

Do users employ a kind of simplified language, e.g. without proper inflection, as they would use towards foreigners without sufficient knowledge of German?

3.3.2. Resources employed

For processing of the FOLKER based transcripts we used the UIMA framework.[7]

The first step is to transform FOLKER format into UIMA based annotations. After this, we initiate a number of linguistic and dialogue based analyses. For these analyses, we used internal and external tools and resources. For example, we integrated resources of GermaNet[8], LIWC (Wolf et al., 2008) and of the project Wortschatz Leipzig. The results of these analyses were exported as XMI documents within UIMA based annotations. These data were exploited for specific research questions realized by specific UIMA consumers.

3.3.3. Hypotheses

We work with the following hypotheses that have to be checked with the transcripts of user companion interactions of N = 130 subjects.

When users start with a linguistic form of a packing/unpacking command that proves to be successful then there will be a strong tendency to continue to use this form.

When users start with a full sentence for a packing/unpacking command then an elliptic version (e.g. 'three jeans') as a follow up is likely as well whereas a telegrammatic version (e.g.'jeans three') should be unlikely.

We interpret the use of full sentences and the use of subsequent elliptic forms as an indication that the resp. subject interprets and accepts the system more like a human dialog partner (especially when politeness particles are used in addition). In contrast we see the use of telegrammatic versions of packing/unpacking commands as indicating a more distant stance towards the system.

Users very likely will not arbitrarily mix telegrammatic versions of packing/unpacking with elliptic versions, but rather stay with the chosen style.

An obvious question: How do these user choices correlate with characteristics of the subjects, e.g. with their socio-

[7] uima.apache.org/

[8] http://www.sfs.uni-tuebingen.de/lsd/

demographic or personality features and/or their computer literacy?

3.3.4. Emotional content

In the experiments reported here we have three sources of utterances with emotional contents:

- self reporting about past emotions in the personalisation phase for all subjects,
- self reporting about current emotions in the intervention phase for the randomly chosen subjects with an intervention and
- spontaneous expression of emotions (e.g. swear words, off talk, self accusations, etc.) especially at the barriers or when problems occur during the interaction.

The self reports of the subjects with intervention are a good indicator for the effectiveness of emotion and affect induction during the last minute phase. From a total of N = 65 subjects with an intervention, only approx. 25 % deny to have experienced unpleasant feelings at the Waiuku barrier. The others explicitly mention to have experienced feelings like anger, disappointment, surprise, stress, time pressure, nervosity.
A detailled linguistic analysis of the various forms of emotional content in the LAST MINUTE transcripts is on the agenda.

4. Discussion and future work

We have presented the current state of the LAST MINUTE corpus. This corpus of recordings from naturalistic interactions between humans and a WoZ simulated companion system excels available corpora with respect to cohort size, volume and quality of data and comes with accompanying data from psychometric questionnaires and from post hoc in depth interviews with participants. The material is a cornerstone for work in the SFB TRR 62 but is as well available for research in affective computing in general.
The long term goal of our joint work is to develop robust classifiers that allow to reliably infer the users' emotional state during the interaction with a companion system thus allowing the companion to appropriately react and to proactively intervene.

Acknowledgment

The presented study is performed in the framework of the Transregional Collaborative Research Centre SFB/TRR 62 "A Companion-Technology for Cognitive Technical Systems" funded by the German Research Foundation (DFG). The responsibility for the content of this paper lies with the authors.
We do thank our students Bastian Kurbjuhn, Friedrich Lüder, Werner Schöne and Stephan Sorgatz for their engaged work as wizards and transcribers and/or help in programming the WoZ software and we express our thanks to Julien Beck, Stefanie Bröter, Andreas Engelhart, Stephan Günther, Viktoria Heine, Sabrina Heising, Albrecht Hultsch, Bianca Kettern, Johanna Kube, Constantin Kwiatkowski, Annette Lederer, Martin Linnemann, Madeleine Luther, Christoph Müller, Sven Plate, Lena Saß, Andreas Schmückert, Sandra Schönfels, Stephan Sorgatz, Natascha Stapf, Dürten Stein, Christian Wackrow, Anita Wilke and Linda Windisch for their transcription and their work in quality assurance of the transcripts.
Special thanks goes to the participants of the Workshop "Linguistische Analysen in der Mensch-Maschine-Interaktion" (Linguistic analyses in human-machine-interaction) in Magdeburg, Oct. 15–16, 2010.

Availability

The LAST MINUTE corpus is available for research purposes upon written request from the authors.

5. References

J. F. Cohn, K. Schmidt, R. Gross, and P. Ekman. 2002. Individual differences in facial expression: Stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings of the International Conference on Multimodal User Interfaces (ICMI 2002)*, pages 491–496.

Roddy Cowie and Marc Schröder. Jan 2005. Piecing together the emotion jigsaw. *Machine Learning for Multimodal Interaction*, pages 305–317.

E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. K. J. Heylen. 2008. The Sensitive Artificial Listener: an induction technique for generating emotionally coloured conversation. In L. Devillers, J-C. Martin, R. Cowie, E. Douglas-Cowie, and A. Batliner, editors, *LREC Workshop on Corpora for Research on Emotion and Affect, Marrakech, Marokko*, pages 1–4, Paris, France. ELRA.

Jörg Frommer, Matthias Haase, Julia Lange, Dietmar Rösner, Rafael Friesen, and Mirko Otto. 2010. Project A3 'prevention of negative dialogue courses' Wizard of Oz experiment operator manual. SFB-Trr-62 working paper, unpublished.

Jörg Frommer, Bernd Michaelis, Dietmar Rösner, Andreas Wendemuth, Rafael Friesen, Matthias Haase, Manuela Kunze, Rico Andrich, Julia Lange, Axel Panning, and Ingo Siegert. 2012. Towards emotion and affect detection in the multimodal LAST MINUTE corpus. In *Eighth international conference on Language Resources and Evaluation (LREC 2012)*. accepted.

Milan Gnjatovic and Dietmar Rösner. 2010. Inducing Genuine Emotions in Simulated Speech-Based Human-Machine Interaction: The NIMITEK Corpus. *IEEE Transactions on Affective Computing*, 1:132–144.

M.A.K Halliday. 1976. A brief sketch of systemic grammar. *System and Function in Language*, pages 3–6.

M. Legát, M. Grůber, and P. Ircing. 2008. Wizard of oz data collection for the czech senior companion dialogue system. In *Fourth International Workshop on Human-Computer Conversation*, pages 1 – 4, University of Sheffield.

G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084, July.

Robert Niese, Ayoub Al-Hamadi, and Bernd Michaelis. 2010. Emotion recognition based on 2d-3d facial feature extraction from color image sequences. *Journal of Multimedia, in print*.

Carl Rogers. 1959. A theory of therapy, personality and interpersonal relationships as developed in the client-centered framework. In S. Koch, editor, *Psychology: A Study of a Science*, volume 3: Formulations of the Person and the Social Context. New York: McGraw Hill.

Dietmar Rösner, Rafael Friesen, Mirko Otto, Julia Lange, Matthias Haase, and Jörg Frommer. 2011. Intentionality in interacting with companion systems – an empirical approach. In Julie Jacko, editor, *Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments*, volume 6763 of *Lecture Notes in Computer Science*, pages 593–602. Springer Berlin / Heidelberg. 10.1007/978-3-642-21616-9_67.

Dietmar Rösner, Jörg Frommer, Rafael Friesen, Matthias Haase, Julia Lange, and Mirko Otto. 2012. LAST MINUTE: a multimodal corpus of speech-based user-companion interactions. In *Eighth international conference on Language Resources and Evaluation (LREC 2012)*. accepted.

Stefan Scherer, Ingo Siegert, Lutz Bigalke, and Sascha Meudt. 2010. Developing an expressive speech labeling tool incorporating the temporal characteristics of emotion. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*.

Thomas Schmidt and Wilfried Schütte. 2010. FOLKER: An Annotation Tool for Efficient Transcription of Natural, Multi-party Interaction. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. AVEC 2011 The First International Audio/Visual Emotion Challenge. In Sidney D Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, volume 6975 of *Lecture Notes in Computer Science*, pages 415–424. Springer Berlin / Heidelberg. 10.1007/978-3-642-24571-8_53.

Margret Selting, Peter Auer, Dagmar Barth-Weingarten, Jörg Bergmann, Pia Bergmann, Karin Birkner, Elizabeth Couper-Kuhlen, Arnulf Deppermann, Peter Gilles, Susanne Günthner, Martin Hartung, Friederike Kern, Christine Mertzlufft, Christian Meyer, Miriam Morek, Frank Oberzaucher, Jörg Peters, Uta Quasthoff, Wilfried Schütte, Anja Stukenbrock, and Susanne Uhmann, 2009. *Gesprächsanalytisches Transkriptionssystem 2 (GAT 2)*. Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion, 10 edition.

2009. Website of the Transregional Collaborative Research Centre SFB/TRR 62. http://www.sfb-trr-62.de/.

Nick Webb, David Benyon, Jay Bradley, Preben Hansen, and Oli Mival. 2010. Wizard of oz experiments for a companion dialogue system: Eliciting companionable conversation. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).

Y. Wilks. 2010. *Close Engagements with Artificial Companions: Key Social, Psychological and Design issues*. John Benjamins, Amsterdam.

Markus Wolf, Andrea B. Horn, Matthias R. Meh, Severin Haug, James W. Pennebraker, and Hans Kordy. 2008. Computergestützte quantitative Textanalyse. volume 54. Hogrefe, Göttingen.

Mining Authors' Personality Traits from Modern Greek Spontaneous Text

Katia Lida Kermanidis
Department of Informatics, Ionian University
7 Tsirigoti Square, 49100 Corfu, Greece
kerman@ionio.gr

Abstract

The present work describes the automatic recognition of authors' personality traits, based on the linguistic properties of their writing. An SVM classifier is applied for the first time to Modern Greek textual features in order to learn the Big Five personality traits for the author. Linguistic features are limited to low-level morphological categorical features. Cross-language findings, even if still premature, are interesting, as several correlations between linguistic properties and personality traits that hold for English, seem to apply in Modern Greek as well. Bootstrapping helps towards improving classification accuracy and avoiding the problem of overfitting.

1. Introduction

The recognition of personality traits is of great importance, as they affect a student's learning patterns and interactive behavior, an employee's performance and willingness to cooperate, a manager's leadership abilities, a teacher's ability to get through and motivate (Mairesse et al., 2007), etc. Personality is traditionally characterized by a set of five features (John and Srivastava, 1999), widely known as the Big Five traits, namely extraversion/introversion, emotional stability/neuroticism, agreeableness/disagreeableness, conscientiousness/unconscientiousness, openness to experience.

A significant research effort is evident nowadays towards the automatic recognition of the big traits, either through the use of language (Mairesse et al., 2006; Mairesse et al., 2007; Mullen, 2011; Mohtasseb and Ahmed, 2009), through the use of speech features (Staiano et al., 2011), or through the use of other media, such as smartphones (Chittaranjan et al., 2011).

Linguistic utterances are strongly connected to the personality of their author. Several research studies have shown that the use of linguistic features, the choice of words and symbols, the statistical properties of a text are indicative of several aspects of the author's personality. Attempts have been made to automatically identify the personality aspects of an author by analyzing his or her text (Mairesse et al., 2006). Their application setting may vary from natural language generation that incorporates individual author stylistic properties (Mairesse and Walker, 2008) to blogger identification (Mohtasseb and Ahmed, 2009) and social meaning extraction (Jurafsky et al., 2009). The text may be a spontaneous monolog (essay) (Mairesse et al., 2007) or chat text (Mullen, 2011). While the former type of text lacks in interactivity and conversational aspects, compared to chat text, it is not restricted in domain and therefore not limited in vocabulary richness.

In the work presented herein, supervised learning is employed for the automatic recognition of the Big Five traits, using elementary (low-level) linguistic features of spontaneously written Modern Greek (MG) essay text. The contribution of the work is two-fold. First, the extent to which personality traits, proven to affect certain linguistic properties in languages like English, also affect MG text properties in the same manner, i.e. the cross-language aspects of the application. To the author's knowledge, this is the first attempt to automatically identify personality traits from MG text. The second goal is to assess how low-level linguistic features can indicate personality tendencies and provide a ranking of the features, based on their impact on the task.

2. Data Collection

The text collection used for the experiments presented herein consisted of spontaneously written monologs (essays). The methodology for creating the essay corpus was similar to the one described in Mairesse et al. (2007): 73 participants (authors: 42 men and 31 women) were asked to write down their thoughts (i.e. what was on their mind at the time of the experiment) spontaneously (without prior notice) for 15-20 minutes. The experiment was repeated ten times, at different time periods, leading to a corpus size of 470 essays. The participants were students of the Department of Informatics of the Ionian University, their ages varying between 18 and 35. The length of the essays varied between 29 and 290 words.

Questionnaires were handed out to every participant in order to establish the ground truth regarding their personality traits. The questionnaire was adopted from John and Srivastava (1999), and consists of 44 5-scale Likkert style questions (1: strongly disagree, 5: strongly agree). Each question represents a personality characteristic that may or may not apply to the participant. Based on the guidelines of John and Srivastava, the participants' answers were translated into Big Five trait values.

Unlike related research that takes into account a wide range of sophisticated features (features denoting cognition, biological processes, relativity etc) (Mairesse et al., 2007; Mohtasseb and Ahmed, 2009). the linguistic features extracted were low-level, in order to enable the

study of the impact of the sophistication level of the features on the task at hand. Only a few psychological markers were taken into account, and no conversational markers (e.g. fillers), as the text consists of monolog essays. Agreement words, for instance, were taken into account for experimental purposes, but were extremely rare (only 8% of the essays contained this type of words). Table 1 contains the exhaustive list of linguistic features and their statistical properties in the corpus. The gender of the author was included to help investigate how it affects the choice of linguistic features, and how it is linked to the Big Five traits.

Morphological and part-of-speech information was obtained using the tool developed by Athens University of Economics and Business (http://nlp.cs.aueb.gr/). The remaining features (positive and negative emotion words, social, agreement, tentative and swear words) were extracted using small manually-crafted lexica.

Feature	Description	Min	Max	Mean	StdDev
Nr of words in essay	essay length	29	290	172.055	58.98
Nr of sentences	number of sentences in the essay	2	19	10.4	4.5
Nr of questions	% of sentences that are questions	0	0.2	0.04	0.073
Elaborated constructions	% of sentences that are complex	0	0.7	0.276	0.271
Positive Emotion Words	% of words denoting positive emotion, e.g. *joy*, *delight* etc.	0	0.054	0.021	0.019
Tentative words	perc. of words like *ίσως* (maybe), *πιστεύω* (I believe that)	0	0.027	0.008	0.007
Social words	% of words denoting inclusion, e.g. *together*, *family*, etc.	0	0.072	0.021	0.022
Agreement words	% of words denoting agreement, e.g. *okay* etc	0	0.007	0.001	0.002
Noun frequency	% of words that are nouns	0.112	0.338	0.163	0.049
Adjective frequency	% of words that are adjectives	0.03	0.124	0.058	0.02
Preposition frequency	% of words that are prepositions	0.025	0.29	0.059	0.037
Article frequency	% of words that are articles	0.003	0.211	0.118	0.052
Pronoun frequency	% of words that are pronouns	0.032	0.158	0.084	0.026
Verb frequency	% of words that are verbs	0.091	0.246	0.15	0.03
Adverb frequency	% of words that are adverbs	0.044	0.163	0.078	0.029
Interjection frequency	% of words that are interjections	0	0.018	0.001	0.003
1st person singular pronouns	number of 1st person singular pronouns	0	6	2.027	1.833
1st person singular pronouns (%)	% of 1st person singular pronouns	0	1	0.167	0.17
Anger words	% of words denoting anger, aggression	0	0.072	0.026	0.027
Swear words	% of words used for 'namecalling', swearing etc.	0	0.013	0.002	0.004
Punctuation marks	% of tokens that are punctuation marks	0.006	0.276	0.114	0.039
Negations	% of words like *όχι* (no), *δεν* (not)	0	0.076	0.025	0.016
Negative Emotion words	% of words denoting negative emotion	0	0.101	0.05	0.031
Present tense verbs	% of verbs that are in the present tense	0.167	1	0.547	0.203
Gender	male or female author				

Table 1: The set of linguistic features.

3. The Big Five Traits

For the automatic classification of the essays the Support Vector Machines classifier implemented in the Weka Machine Learning Workbench (www.cs.waikato.ac.nz/ml/weka/) was used, because it was shown to achieve better performance in previous work (Mairesse et al., 2007). It employs the Sequential Minimal Optimization algorithm for training and parameters were set to a second degree polynomial kernel function and a complexity parameter with value 2.0. Experiments were run using 10-fold cross validation.

Due to the small size of the dataset, metalearning was also experimented with in order to increase stability and classification accuracy and to avoid overfitting. Bootstrap aggregating (Bagging) was employed (Breiman, 1996), i.e. the SVM classifier was run ten times, each time on a different 60% part of the dataset. The majority vote of the ten models determines the final prediction. Classification performance is shown in Tables 2 and 3 for all Big Five traits and both class values

(Y:yes, N:no).

Even though the differences between the stand-alone and the meta-learning results are not statistically significant at the 0.05 level, as was shown by paired t-testing, it is clear that bagging tends to improve classification performance. It is interesting to observe that precision and recall for the positive class is higher in all cases compared to the negative class. To a large part, this can be attributed to the nature of the questionnaire, and the algorithm that derives the trait label from it (John and Srivastava, 1999). The number of questions that are more straightforwardly linked to the positive values of the five traits is larger than the one linked to the negative values.

Trait	SVMs – stand alone			
	Precision		Recall	
	Y	N	Y	N
Extraversion	0.67	0.59	0.65	0.61
Emotional Stability	0.62	0.40	0.73	0.29
Agreeableness	0.69	0.50	0.76	0.41
Conscientiousness	0.57	0.21	0.60	0.19
Openness	0.80	0.50	0.79	0.52

Table 2: Classification Performance with SVMs as a stand-alone classifier.

Trait	SVMs – Bagging			
	Precision		Recall	
	Y	N	Y	N
Extraversion	0.67	0.59	0.65	0.61
Emotional Stability	0.65	0.44	0.69	0.39
Agreeableness	0.71	0.52	0.74	0.48
Conscientiousness	0.64	0.33	0.75	0.23
Openness	0.79	0.56	0.86	0.43

Table 3: Classification Performance with SVMs in a bagging meta-learning schema.

Direct comparison of the achieved performance with previous approaches is not meaningful due to the significant differences in the employed feature sets. Mairesse et al. (2007) report a classification accuracy ranging between 54.9% (extraversion) and 62.1% (openness) with the same learning algorithm and a similar essay corpus. Their feature set, however, is a combination of the LIWC (Pennebaker et al., 2001) and the MRC (Coltheart, 1981) features, i.e. much more extensive and higher-level than the one used in the present work.

4. Conclusion

Even though the results for conscientiousness are low, it is interesting that the features reported in the literature to affect this trait (negations), constitute the most important feature for classifying essays regarding this trait in the experiments described herein as well. The lack of fillers in the present feature set, used in conversational text and proven to indicate conscientiousness, might account partly for the low accuracy. Questions, negations, words of anger and swearing, and the use of adverbs seem to be indicative of extraversion. According to the results, frequent use of adjectives, agreement words and interjections are indicative of emotional stability. The frequent use of interjections and avoidance of first person pronouns are indicators of agreeableness, while more adjectives and less adverbs indicate openness to experience. These remarks have been derived by performing feature selection experiments based on the Information Gain value of the features.

Another interesting remark is that women tend to use more elaborate constructions than men, while the latter use more swearwords. Also, the frequent use of swearwords seems to be positively correlated with men's extraversion. Naturally, these are only indications, derived from a particular experimental setting, interesting for experimentation, but hardly a basis to draw concrete conclusions from.

The results support the claim that the link between linguistic expression and personality traits remains more or less the same across languages. The use of conversational text would probably enable the study of other linguistic features, and reveal other dependencies with personality traits, and is a future research direction to be explored. However, conversational text alone lacks significant features, present in essay text (e.g. elaborate constructions), which has been proven in previous as well as the present work to constitute a significant indicator of the author.

It is evident, in any case, that the use of even low-level linguistic features is indicative of the writer and his style.

5. Acknowledgements

The author would like to thank Paraskevi Pasxali for her help in collecting the data, and setting up the dataset.

6. References

Breiman, L. (1996). Bagging Predictors. *Machine Learning*, 24(2), pp. 123–140.

Chittaranjan, G., Blom, J. and Gatica-Perez, D. (2011). Who's Who With Big Five: Analyzing and Classifying Personality Traits With Smartphones. Proceedings of the 15th IEEE Annual International Symposium on Wearable Computers, San Francisco.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, 33A, pp. 497–505

John, O. P. and Srivastava, S. (1999). The Big Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. *Handbook of Personality: Theory and Research*. Guilford, New York.

Jurafsky, D. et al. (2009). Extracting Social Meaning: Identifying Interactional Style in Spoken

Conversation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (NAACL HLT).

Mairesse, F. and Walker, M. (2008). Trainable Generation of Big Five Personality Styles Through Data-driven Parameter Estimation. *Proceedings of the Conference of the Association of Computational Linguistics* (ACL), pp. 165-173.

Mairesse, F. et al. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. *Journal of Artificial Intelligence Research*, 30, pp. 457-501.

Mairesse, F. and Walker, M. (2006). Automatic Recognition of Personality in Conversation. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (NAACL HLT).

Mohtasseb, H. and Ahmed, A. (2009). Mining Online Diaries for Blogger Identification. Proceedings of the *International Conference on Data Mining and Knowledge Engineering*, pp. 295-302.

Mullen, J. (2011). The Impact of Computer Use on Employee Performance in High-Trust Professions: Re-examining Selection Criteria in the Internet Age. *Journal of Applied Social Psychology*, 41(8), pp. 2009-2043.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Inquiry and Word Count: LIWC 2001*. Lawrence Erlbaum, Mahwah, NJ.

Staiano, J. et al., (2011). Automatic Modeling of Personality States In Small Group Interactions. *Proceedings of the 19th ACM International Conference on Multimedia*, pp. 989-992.

Building Corpora for Figurative Language Processing: The Case of Irony Detection

Antonio Reyes[1] [2], **Paolo Rosso**[2]

[1] Language Technology Lab
Instituto Superior de Intérpretes y Traductores, Mexico
antonioreyes@isit.edu.mx
[2] Natural Language Engineering Lab — ELiRF
Universitat Politècnica de València, Spain
prosso@dsic.upv.es

Abstract

Figurative language is one of the most arduous topics that natural language processing (NLP) has to face. Unlike literal language, the former takes advantage of linguistic devices, such as metaphor, analogy, ambiguity, irony, sarcasm, and so on, in order to communicate more complex meanings, which usually represent a serious problem, not only for computers, but for humans as well. In this article we describe the problem of figurative language processing concerning corpus-based approaches. This type of language is quite common in web contents; however, its automatic processing entails a huge challenge, both theoretically as pragmatically. Here we describe the need of automatically building training corpora with objective and reliable data. In this respect, we are focused on addressing a quite complex device: irony. Such linguistic phenomenon, which is widespread in web content, has important implications for tasks such as sentiment analysis, opinion mining, or even advertising.

1. Introduction

Language, in all its forms, is the most natural and important mean of conveying information. However, given its social nature, it cannot be conceptualized only in terms of grammatical issues. In this respect, while it is true that grammar regulates language in order to have a non-chaotic system, it is also true that language is dynamic, and accordingly, a live entity. This means that language is not static, rather it is in constant interaction between the rules of its grammar and its pragmatic use. For instance, the idiom *"all of a sudden"* has a grammatical structure which is not made intelligible only by knowledge of the familiar rules of its grammar (Fillmore et al., 1988), but by inferring pragmatic information as well. This latter provides the knowledge that, in the end, gives sense to the idiom.

Emphasizing the social aspect of language, modern linguists deem language as a continuum of symbolic structures in which lexicon, morphology, and syntax form a continuum which differs along various parameters but can be divided into separate components only arbitrarily (Langacker, 1991). Language, thus, is viewed as an entity whose components and levels of analysis cannot be independent nor isolated. On the contrary, they are embedded in a global system that depends on cognitive, experiential, and social contexts, which go far beyond the linguistic system proper (Kemmer, 2010). Let us consider the following example:

1. "I really need some antifreeze in me on cold days like this".

Example 1 is fully understandable only within a context in which the sense is given by figuring out the analogy between $antifreeze$ (referential knowledge: antifreeze is a liquid) and liquor (inferential knowledge: antifreeze is a liquid, liquor is a liquid, antifreeze is a liquor).

In this context, the following sections introduce the theoretical background concerning figurative language (Section 2.), describe the problem of dealing with figurative language in a technological framework (Section 3.), report on how we approach the task of automatically building a training corpus for figurative language processing (Section 4.), and conclude with some final remarks about our approach and its further implications (Section 5.).

2. Literal and Figurative Language

Traditionally, language has been described from dichotomous points of view: *langue* $vs.$ *parole*, signifier $vs.$ signified, synchrony $vs.$ diachrony, paradigmatic $vs.$ syntagmatic, oral $vs.$ written, an so on. In this section, another dichotomy will be discussed: literal language $vs.$ figurative language. The simplest definition of literal language is related to the notion of $true$, $exact$ or $real$ meaning; i.e. a word (isolated or within a context) conveys one single meaning (the one conventionally accepted), which cannot be deviated. In Saussure's terms, literal meaning is corresponded with a perfect dichotomy of signifier and signified (cf. (de Saussure, 1974)). Some experts, in addition, have noticed certain properties of literal meaning: it is direct, grammatically specified, sentential, necessary, and context-free (see (Katz, 1980; Searle, 1978; Dascal, 1987)). Hence, it is assumed that it must be invariant in all contexts. For instance, the word $flower$ can only refer to the concept of plant, regardless of its use in different communicative acts or discourses (e.g. botany, evolution, poetry).

On the other hand, figurative language could be regarded as the simple oppositeness of literal language. Thus, whereas the latter is assumed to communicate a direct meaning, the former is more related to the notion of conveying indirect or veiled meanings. For instance, the word $flower$, which literally refers only to the concept of plant, speaking figu-

ratively can refer to several concepts, which not necessarily are linked to plants. Thereby, it can be used instead of concepts such as beauty, peace, purity, life, and so on, in such a way its literal meaning is *intentionally* deviated in favor of secondary interpretations[1].

Although, at first glance, this distinction seems to be clear and sufficient on its own, figurative language involves basic cognitive processes rather than only deviant usage (Peters, 2004). Therefore, it is necessary going deeper into the mechanisms and processes that differentiate both types of languages.

In accordance with classical perspectives, the notions of literalness and figurativity are viewed as pertaining directly to language; i.e. words have literal meanings, and can be used figuratively (Katz, 1980; Searle, 1978; Dascal, 1987). Figurative language, therefore, could be regarded as a type of language that is based on literal meaning, but is disconnected from what people learn about the world [or about the words] based on it [them] (Bergen, 2005). Thus, by breaking this link, literal meaning loses its primary referent and, accordingly, the interpretation process becomes senseless. Let us consider Chomsky's famous example to explain this issue:

2. "Colorless green ideas sleep furiously" (Chomsky, 1957).

Beyond grammatical aspects, in example 2 is possible to observe that, either phonologically or orthographically, Chomsky's example is fully understandable in terms of its linguistic constituents[2]. However, when interpreting such constituents in context, its literal meaning is completely nonsensical. For instance, the bigrams [colorless green] or [green ideas] are sufficiently disconnected from their conventional referents for being able to produce a coherent interpretation. Thus, in order to make the example understandable, secondary interpretations are needed. If such interpretations are successfully activated, then figurative meaning is triggered[3] and, accordingly, a more coherent interpretation can be achieved. Based on this explanation, literal meaning could be deemed as denotative, whereas figurative meaning, connotative; i.e. figurative meaning is not given a priori, rather, must be implicated. Furthermore, in (Lönneker-Rodman and Narayanan, 2008), authors point out that figurative language can tap into conceptual and linguistic knowledge (as in the case of idioms, metaphor, and some metonymies), as well as evoke pragmatic factors in interpretation (as in indirect speech acts, humor, irony, or sarcasm). In accordance with the assumptions, an expected conclusion is to conceive the processes of interpreting figurative language much more complex than the ones performed when interpreting literal language.

[1] It is worth noting that such secondary interpretations are not guaranteed. Their success will depend on several factors, both linguistic as extra-linguistic.

[2] It is worth stressing that this sentence is an intentional example of semantic senseless, whose meaning (either literal or figurative) is supposed to not exist. However, here is used to precisely exemplify the nonsensical effect produced by figurative contents. Most of them, finally, are senseless on their own, and need a pragmatic anchor to correctly interpret their meanings.

[3] According to (Sikos et al., 2008), understanding figurative language often involves an interpretive adjustment to individual words; i.e. not all the constituents of the example trigger a figurative meaning on their own, rather, this is usually triggered by manipulating individual words.

3. Figurative Language and Web Content

Web-based technologies have become a significant source of data in a variety of scientific and humanistic fields. Such technologies provide a rich vein of information that is easily mined. User-generated content (such as text, audio and images) provides knowledge that is topical, task-specific, and dynamically updated to broadly reflect changing trends, behavior patterns and social preferences. In this context, figurative language can be found on almost every web site in a variety of guises and with varying degrees of obviousness. For instance, when analyzing instances of irony, one of the most important micro-blogger sites: Twitter, allows its users to self annotate their posts with user-generated tags (or hashtags according to Twitter's terminology). Thus, the hashtag #irony is used by people in order to self-annotate all varieties of irony, whether they are chiefly the results of deliberate word-play or merely observations of the humor inherent in everyday situations (e.g. "Sitting in the eye-doctor's office, waiting for the doctor to see me"), or simply sarcastic expressions (e.g. "I thank God that you are unique!").

3.1. The Core of the Problem

Although the arguments given in the previous sections provide some elements to determine what figurative language is, a major question still remains: how to differentiate between literal language and figurative language (theoretically and automatically)? The examples given so far have shown some of their main characteristics; however, based on that information, there is not way of totally affirming that example 1 is more figurative than example 2. Finally, both examples could be expressing, either of literal or figurative language. To be able to provide arguments for differentiating both linguistic realities, a crucial extra-linguistic element (with linguistic repercussion) must be highlighted: **intentionality**. Beyond mechanisms to explain why figurative language requires much more cognitive efforts to correctly interpret its meaning, the most important issue is that the previous examples are simply sequences of words with semantic meaning. Perhaps, such meaning is very clear (literalness), or perhaps is senseless (figurativity), but they could be explained in terms of performance and competence or even as a matter of correctness . However, such difference could be motivated by the need of maximizing a communicative success (cf. (Sperber and Wilson, 2002)). Such need would be then the element that will determine what type of information has to be profiled. If a literal meaning is profiled, then certain intention will permeate the statement. This intention will find a linguistic repercussion by selecting some words or syntactic structures to successfully communicate what is intended. In contrast, if the figurative meaning is profiled, then the intention will guide the choice of others elements to ensure the right transmission of its content. It is likely that such content cannot be ac-

complished, but in this case, the failure will not lay on the speaker's intention, rather, on the hearer's skills to interpret what is communicated figuratively. Let us observe the following sentences to clarify this point.

3. "The rainbow is an arc of colored light in the sky caused by refraction of the sun's rays by rain" (cf. WordNet (Miller, 1995) v. 3.0).

4. "The rainbow is a promise in the sky".

Whereas in example 3 the intention is to describe what a rainbow is, in example 4 the intention is to communicate a veiled meaning, motivated and understandable by a specific conceptual context. In each statement the speaker has a communicative need, which is solved by maximizing certain elements. Thus, in the first example, the communicative success is based on making a precise description of a rainbow (note that all the words in this context are very clear in terms of their semantic meaning), whereas in the second, is based on deliberately selecting elements that entail secondary and nonliteral relations: [rainbow - promise], [promise - sky].

4. User-Generated Tags: Explicit Intentionality

Once argued that intentionality is one of the most important mechanisms to differentiate literal from figurative language, it is worth noting that user-generated tags provide specific elements to deliberately express different types of figurative contents: metaphor, allegory, irony, similes, analogy, and so on. In this respect, we are focused on the case of irony.

Irony (and most figurative language) is very subjective and often depends on personal appreciation[4]. Therefore, the task of collecting ironic examples (positive data) is quite challenging. In addition, as noted in (Reyes and Rosso, 2011; Reyes et al., 2012), the boundaries to differentiate the different types of irony (mostly verbal irony and situational irony) are very fuzzy indeed: non-expert people usually use an intuitive and unspoken definition of irony rather than one sanctioned by a dictionary or a text-book. Hence, such task becomes any harder.

4.1. A Basic Sample

Although a manual annotation is supposed to be the best way of obtaining reliable information in corpus-based approaches, in tasks like this one, such approach is hard to be achieved. First, there are not formal elements to accurately determine the necessary components to label any text as ironic. Then, in the case that we had a prototype of ironic expressions, its discovery is a time-consuming manual task[5]. Finally, linguistic competence, personal appreciation, moods, and so on, make irony quite subjective; therefore, any annotation agreement faces the complexity of standardizing annotation criteria. That is why we decided to use examples labeled with user-generated tags, which are **intentionally** focused on particular topics[6]. By opting for this approach, we eliminate the inconveniences above mentioned: such examples are self-annotated (thus, it is not necessary the presence of "human annotators" to manually (and subjectively) collect and label positive examples). In addition, positive examples can be retrieved effortless taking advantage of their tags (thus, it is likely having thousands of examples in a short time).

In this context, we here describe how we have taken advantage of the user-generated tags in order to build a training corpus for the irony detection task. To this end, we are focused on one of the current trendsetters in social media: the Twitter micro-blogging service. We first determine a membership criterion for including a tweet in the corpus: each should contain a specific $hashtag$ (i.e. the user-generated tag according to Twitter's terminology). The hashtags selected are #irony, in which a tweet explicitly declares its ironic nature, as well as #education, #humor, and #politics, to provide a large sample of potentially non-ironic tweets. These hashtags are selected because when using the #irony hashtag, people employ (or suggest) a family-resemblance model of what it means (cognitively and socially) for a text to be ironic. In this respect, a text so-tagged may not actually be ironic by any dictionary definition of irony, but the tag reflects a tacit belief about what constitutes irony. Based on these criteria, we collect a training corpus of 40,000 tweets, which is divided into four parts, comprising one self-described positive set and three other sets that are not so tagged, and thus assumed to be negative. The final corpus contains 10,000 ironic tweets and 30,000 largely non-ironic tweets. Some statistics are given in Table 1. It is worth noting that all the hashtags were removed. No further preprocessing was applied at this point.

Table 1: Statistics in terms of tokens per set.

	#irony	#education	#humor	#politics
Vocabulary	147,671	138,056	151,050	141,680
Nouns	54,738	52,024	53,308	57,550
Adjectives	9,964	7,750	10,206	6,773
Verbs	29,034	18,097	21,964	16,439
Adverbs	9,064	3,719	6,543	4,669

Due to the intrinsic characteristics concerning writing habits in technological platforms such as blogs, cell phones, etc., it is very likely the presence of many errors in the documents, as well as the presence of duplicate documents, or even pointless information. In order to minimize such errors, several measures can be applied. Here we outline just one of them: the Jaccard distance. Such metric measures the dissimilarity between two samples, and is calculated according to Formula 1. The Jaccard distance is here used to estimate the overlap between the ironic set and each of the three non-ironic ones. In addition, it should help mini-

[4]That is why the importance of considering both linguistic as paralinguistic features when modelling this complex device.

[5]According to (Peters and Wilks, 2003), this is a reason for the restricted number of attested instances of figurative language in texts. In addition, it is worth noting that irony appears quite often in discourse. For instance, in (Carvalho et al., 2011), authors indicate that irony is present in approximately 11% of their data.

[6]Recall the role of intentionality in the process of communicating the figurative intent.

mizing the likelihood of noise arising from the presence of typos, common misspellings, and the abbreviations that are endemic to short texts.

$$J_\delta(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (1)$$

Results in Table 2 suggest a significant difference between the vocabularies of the four tweet sets. As one might expect, this difference is least pronounced between sets #irony and #humor. After all, irony is most often used to communicate a humorous attitude or insight, as in examples 5 and 6, in which the tweet was tagged as #irony:

Table 2: Jaccard distance among sets.

	J_δ(A,B)
(irony, education)	0.8233
(irony, humor)	0.8565
(irony, politics)	0.8246

5. Just think: every time I breathe a man dies. —A friend: Have you tried to do something about bad breath?

6. I find it humorously hypocritical that Jeep advertises on TV about how we shouldn't watch tv in favor of driving their vehicles.

Finally, it is worth noting that this approach is useful to the spread of researches related to figurative language, as well as to palliate the lack of resources for figurative language processing, and especially, to face tasks in which the scarcity of data, the subjectivity of the task, or the impossibility of making personal interviews, are challenges to be tackled[7].

5. Final Remarks

In this article we have discussed the problem of figurative language and its automatic processing. In particular, we were focused on addressing the task of automatically building training corpora when facing one of the most complex figurative devices: irony. Although the approach here described is slightly theoretical, it has important implications for tasks such as sentiment analysis (cf. (Reyes et al., 2012) about the importance of determining the presence of irony in order to assign fine-grained polarity levels), trend discovery (cf. (Reyes and Rosso, 2011; Reyes and Rosso, In press), where authors note the impact of user-generated tags for discovering people's trends in ironic documents), or opinion mining (cf. (Sarmento et al., 2009), about the role of irony in discriminating negative from positive opinions). In the future, we plan to approach irony detection from each of its angles building corpora that could consider also valuable information such as gestural information, tone, paralinguistic cues, etc. (cf. (Cornejol et al., 2007)).

[7]The relevance of approaches like this one can be confronted in (Reyes and Rosso, 2011): in such work authors collected a corpus for irony detection only with reviews posted in Amazon.

Last but not least, it would be also interesting try to model irony taking into consideration the visual stimulus of brains responses when people have to process ironic statements (cf. (Mars et al., 2008)).

Acknowledgements

This work has been done in the framework of the VLC/CAMPUS Microcluster on Multimodal Interaction in Intelligent Systems and it has been partially funded by the European Commission as part of the WIQEI IRSES project (grant no. 269180) within the FP 7 Marie Curie People Framework, and by MICINN as part of the Text-Enterprise 2.0 project (TIN2009-13391-C04-03) within the Plan I+D+I.

6. References

B. Bergen. 2005. Mental Simulation in Literal and Figurative Language Understanding. In Seana Coulson, editor, *The Literal And Nonliteral in Language and Thought*, pages 255–280. Peter Lang Publishing, September.

P. Carvalho, L. Sarmento, J. Teixeira, and M. Silva. 2011. Liars and saviors in a sentiment annotated corpus of comments to political debates. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, pages 564–568.

N. Chomsky. 1957. *Syntactic Structures*. Mouton and Co, The Hague.

C. Cornejol, F. Simonetti, N. Aldunate, A. Ibáñez, V. López, and L. Melloni. 2007. Electrophysiological evidence of different interpretative strategies in irony comprehension. *Journal of Psycholinguist Research*, 36:411–430.

M. Dascal. 1987. Defending literal meaning. *Cognitive Science*, 11(3):259–281.

F. de Saussure. 1974. *Course in general linguistics*. Fontana, London.

C. Fillmore, P. Kay, and M. O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

J. Katz. 1980. *Propositional structure and illocutionary force: A study of the contribution of sentence meaning to speech acts*. Harvard University Press.

S. Kemmer. 2010. About cognitive linguistics: Historical background. `http://www.cognitivelinguistics.org/cl.shtml`. Online on August 25, 2011.

R. Langacker. 1991. *Concept, Image and Symbol. The Cognitive Basis of Grammar*. Mounton de Gruyter.

B. Lönneker-Rodman and S. Narayanan. 2008. Computational approaches to figurative language.

R. Mars, B. Rogier, S. Debener, T. Gladwin, L. Harrison, P. Haggard, J. Rothwell, and S. Bestmann. 2008. Trial-by-Trial Fluctuations in the Event-Related Electroencephalogram Reflect Dynamic Changes in the Degree of Surprise. *J. Neurosci.*, 28(47):12539–12545.

G. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

W. Peters and Y. Wilks. 2003. Data-driven detection of figurative language use in electronic language resources. *Metaphor and Symbol*, 18(3):161–173.

W. Peters. 2004. *Detection and Characterization of Figurative Language Use in WordNet*. Ph.D. thesis, University of Sheffield, Sheffield, England.

A. Reyes and P. Rosso. 2011. Mining subjective knowledge from customer reviews: A specific case of irony detection. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 118–124.

A. Reyes and P. Rosso. In press. Making objective decisions from subjective data: Detecting irony in customers reviews. *Decision Support Systems*.

A. Reyes, P. Rosso, and D. Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data and Knowledge Engineering*. DOI: 10.1016/j.datak.2012.02.005 `http://dx.doi.org/10.1016/j.datak.2012.02.005`.

L. Sarmento, P. Carvalho, M. Silva, and E. de Oliveira. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29–36.

J. R. Searle. 1978. Literal meaning. *Erkenntnis*, 13(1):207 – 224.

L. Sikos, S. Windisch Brown, A. Kim, L. Michaelis, and M. Palmer. 2008. Figurative language: “meaning" is often more than just a sum of the parts. In *Proceedings of the AAAI 2008 Fall Symposium on Biologically Inspired Cognitive Architectures.*, pages 180–185.

D. Sperber and D. Wilson. 2002. Relevance theory. *Handbook of Pragmatics*, 42(5):607–632.

Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives

Marcela Charfuelan, Marc Schröder

Language Technology Laboratory, DFKI GmbH
Berlin D-10559 and Saarbrücken D-66123, Germany
{marcela.charfuelan,marc.schroeder}@dfki.de

Abstract

We investigate possible correlations between sentiment analysis scores obtained for sentences of Mark Twain's novel "The Adventures of Tom Sawyer" and acoustic features extracted from the same sentences in the corresponding audiobook. We have found that scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean fundamental frequency (F0). We have designed an experiment intended to predict the levels of acoustic expressivity in arbitrary text using sentiment analysis scores and the number of words in the text.

1. Introduction

In this paper we investigate possible correlations between sentiment analysis scores obtained for sentences of Mark Twain's novel "The Adventures of Tom Sawyer" and acoustic features extracted from the same sentences in the corresponding audiobook. In the audiobook a single speaker reads the whole novel, the narration is lively and expressive and the speaker impersonates or performs several characters apart from the narrator himself.

From a theoretical point of view, narratives have been studied as a context for the integration of language and emotion. According to (Reilly and Seibert, 2003) and the references in this work, evaluative information in narratives can be conveyed/packaged in several ways: "lexically", for example, using intensifiers, modals or hedges to reflect speaker attitude; "syntactically" as in relative clauses, which commonly function as asides to comment on a person's behaviour/character; and "paralinguistically", by emotional facial expression, gesture and affective prosody that can effectively convey narrator attitude or reflect the inferred emotions of a character.

Due to the lively character of narration in audiobooks, these have been recently used in several studies related to clustering of expressive speech styles (Székely et al., 2011), expressivenes of speech (Wang et al., 2006) or automatic selection of diverse speech corpora for improving automatic speech synthesis (Braunschweiler et al., 2011a). Audiobooks might help to tackle some of the nowadays key problems on speech synthesis technology: unlabelled prosodic and voice quality variations; expressive speech; large corpora of non-studio-quality speech (Blizzard Challenge, 2012). At the same time audiobooks might also contribute to simplify some of the most difficult problems to progress with synthesis from social signalling corpora: lack of phonetic coverage, lack of single-user speech, and lack of textual transcriptions.

In this paper first we describe the data analysed in Section 2., then in Sections 3. and 4. we describe the sentiment scores obtained for sentences in the book and the acoustic features extracted from the corresponding audio data. In Section 5. we describe two experiments intended to investigate the possible correlation of the previous scores and features and the possibility of using sentiment scores from arbitrary text to predict an acoustic level of "expressivity". Preliminary results and future work are presented in Section 6.

2. Data

The data analysed is the audiobook "The adventures of Tom Sawyer" available at LibriVox (LibriVox, 2012) and its associated text available in Project Gutenberg (Project Gutenberg, 2012). The audiobook has been split into prosodic phrase level chunks, which corresponds to the sentences analysed in this work. The sentence segmentation and orthographic text alignment of the audiobooks has been performed using an automatic sentence alignment method - LightlySupervised - described in (Braunschweiler et al., 2011b). The number of sentences analysed is 5119 corresponding to 17 chapters and approximately 6.6 hours of recordings at 44100 Hz. The books were read by John Greeman, an American English narrator.

3. Sentiment scores

The sentiment scores were obtained in two steps. First, summary statistical information about individual words was extracted using the data and methods of (Potts and Schwarz, 2010) and (Potts, 2011a). Second, to combine these word-level scores effectively in order to make predictions about full sentences, a maximum entropy classifier was trained on a large, diverse collection of texts from social media sources. The reader is referred to these publications for more details about the system as well as (Potts, 2011b) for data and available resources. In the following we summarise the sentiment scores used in this study:

- Scores derived from IMDB reviews using machine learning techniques (Bo et al., 2002):
  - ImdbEmphasis: a sentiment score for emphasis vs. attenuating
  - ImdbPolarity: a sentiment score for positive vs. negative
- OpinionLexicon: sentiment scores by lexicon lookup using Bing Liu's lexicon, which is a list of positive

and negative opinion words or sentiment words for English (around 6800 words) that has been compiled over many years (Liu, 2011).

- SentiWordnet (Wordnet entries with added sentiment scores) negative and positive value:
  - SentiWordNetNeg
  - SentiWordNetPos
- Scores derived from the Experience Project: this project is a social networking website that allows users to share stories about their own personal experiences, users write typically very emotional stories about themselves, and readers can then chose from among five reaction categories to the story (Potts, 2011b). Data from this project has been used to derive the following reaction scores:
  - Hugs: Sympathy reader reaction score
  - Rock: Positive-exclamative reader reaction score.
  - Teehee: Amused/light-hearted reader reaction score.
  - Understand: Solidarity reader reaction score.
  - Wow: Negative-exclamative reader reaction score.
- Predicted negative (Neg) and positive (Pos) probability derived by training a model with the previous scores:
  - Neg, Pos
  - Polar: calculated as Pos-Neg, this is a kind of predicted polarization score, examples of very positive and very negative polarity scores are presented in Table 1.

Text	Polar
Well, goodness gracious!	1.00
Luck!	1.00
I love thee well!	1.00
Glory was sufficient.	0.99
Tom's astonishment was boundless!	0.99
Good!	0.99
...	
Kill?	-1.00
It's awful.	-1.00
Hateful, hateful, hateful!	-1.00
Crash!	-1.00
Bother!	-1.00
It's that dreadful murder.	-1.00

Table 1: Text examples of very positive and very negative polarity scores.

4. Acoustic features

We have extracted well known acoustic correlates of emotional speech: mainly prosody or fundamental frequency (F0) related features, some intonation related measures (F0 contour measures) and voicing strengths features, that have been used to model and improve excitation in vocoded speech. The following features and measures have been calculated:

- F0 and F0 statistics, mean, maximum, minimum and range. F0 values were extracted with the snack tool (Sjölander, 2012).
- Duration in seconds per sentence.
- Average energy, calculated as the short term energy ($\sum s^2$) averaged by the duration of the sentence in seconds.
- Number of voiced frames, number of unvoiced frames and voicing rate calculated as the number of voiced frames per time unit.
- F0 contours, as in (Busso et al., 2009) we have extracted slope (a1), curvature (b2) and inflexion (c3); these measures are estimated by fitting a first-, second- and third-order polynomial to the voiced F0 values extracted from each sentence:

$$y = a_1 * x + a_0 \quad (1)$$

$$y = b_2 * x^2 + b_1 * x + b_0 \quad (2)$$

$$y = c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \quad (3)$$

- Voicing strengths estimated with peak normalised cross correlation of the input signal (Chu, 2003). The correlation coefficient for a delay t is defined by :

$$c_t = \frac{\sum_{n=0}^{N-1} s(n)s(n+1)}{\sqrt{\sum_{n=0}^{N-1} s^2(n) \sum_{n=0}^{N-1} s^2(n+t)}} \quad (4)$$

  Five bandpass voicing strengths are calculated, that is, the input signal is filtered into five frequency bands; mean statistics of this measure are extracted.

5. Experiments

5.1. Correlation analysis

Pairwise correlation between the previously described sentiment scores and acoustic features was performed. We have found correlations mainly between average energy and mean F0 and sentiment scores derived from IMDB reviews and reader reaction scores. Table 2 shows the higher correlation values between these scores and features. The correlation with other sentiment features was very low, in particular no correlation at all was found between F0 contour features and sentiment scores. These results also show that the sentiment scores that come from lexicons are not correlated at all with acoustic features, whereas scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative.

	Acoustic features	
Sentiment scores	Energy	mean_F0
ImdbEmphasis	0.51	0.38
ImdbPolarity	-0.33	-0.31
Teehee	0.29	0.13
Wow	-0.17	-0.30
Polar	-0.13	-0.14

Table 2: Pairwise correlation between sentiment scores and acoustic features.

5.2. Predicting "expressivity"

In a further experiment we investigate if we can predict some measure of "expressivity" just on the basis of sentiment scores. Our measure of expressivity is the first principal component value (PC1) after a principal component analysis (PCA) of all the acoustic features extracted from the data. A PC1 value per sentence was calculated, and we have empirically found that positive values of PC1 most of the time correspond to sentences of the narrator in a more or less neutral voice, and negative values most of the time correspond to expressive sentences where the speaker impersonates one of the characters in the book (childish voice, women voice. etc.). To corroborate this, we have manually annotated the first two chapters of the book according to narrator and the characters the speaker performs. Figure 1 shows the variation of mean F0, ImdbEmphasis and PC1 per sentence in chapter 01, for the narrator and other impersonated characters. In this Figure we can also observe that the values for "other" characters present higher excursion than for the "narrator".

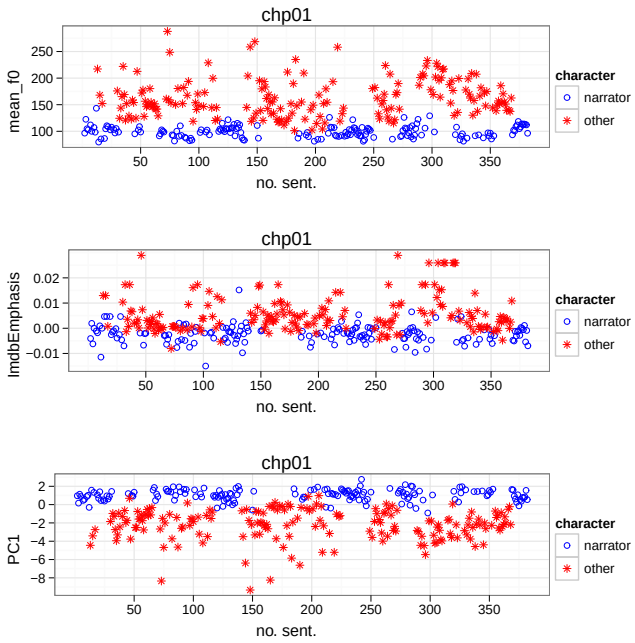

Figure 1: Mean F0, ImdbEmphasis scores and PC1 values for the sentences in chapter 01 of "The Adventures of Tom Sawyer".

Multiple linear regression (MLR) of sentiment scores (plus number-of-words) was used to train a prediction model of the acoustic PC1 feature; sequential floating forward selection (SFFS) was used to find the best sentiment score predictors. Statistical analysis, MLR and SFFS, were performed with R (R Development Core Team, 2009). We have found that the model fits well the training data. Figure 2 shows in blue the PC1 values obtained per sentence for chapter 02 of the book; the predicted values are indicated in red and the prediction error in black. Averaging the results obtained for every chapter, we have found that PC1 is predicted with a prediction error of 1.21 when using just sentiment features; the prediction error improves to 0.62 when using number-of-words in the sentence as another predictor feature.

To evaluate how well the model can predict a level of expressivity with unseen data, we used the annotated chapters 01 and 02 as test data and the rest of the data to train a model. For training a predictor model of PC1 we used all the acoustic features presented en Section 4.; the learnt parameters after the SFFS multiple linear regression are:

$$\begin{aligned} PC1 = &-1.64 + 0.12 \times num\_words\_sentence \\ &- 48.0 \times ImdbEmphasis + 11.3 \times ImdbPolarity \\ &+ 2.24 \times SentiWordNetNeg - 1.78 \times Teehee \\ &- 3.66 \times Understand - 1.17 \times OpinionLexicon \\ &+ 0.6 \times Hugs + 0.44 \times SentiWordNetPos \end{aligned} \quad (5)$$

Using this equation a PC1 value is predicted for the utterances of chapters 01 and 02, the value is further used to determine whether the utterance is character type "narrator" (predicted PC1 $>= 0$) or "other" (predicted PC1 < 0). Since we have character annotations of these two chapters we can compare the annotated character and the predicted one. The character prediction results for 345 utterances of chapter 01 and 271 utterances of chapter 02 are presented in Table 3. Examples of utterances predicted as "narrator" and "other" in chapter 01 are presented in Table 4.

	Chapter 01		Chapter 02	
Character	Narrator	Other	Narrator	Other
Narrator	79.8	30.1	92.0	34.0
Other	20.2	69.9	8.0	66.0
Diagonal	73.3%		81.5%	

Table 3: Character prediction for chapters 01 and 02 using number of word, sentiment scores and the learnt model in equation 5.

We can observe in Table 3 that the character types in chapter 02 were better predicted than in chapter 01. Two observations might explain why "expressivity" in chapter 01 was more difficult to predict: first, the PC1 values of chapter 01 present higher excursion than chapter 02 and second the sentences in chapter 01 are shorter in average than in chapter 02. Chapter 01 has 12.3 words in average per sentence (minimum 1 and maximum 80 words) and chap-

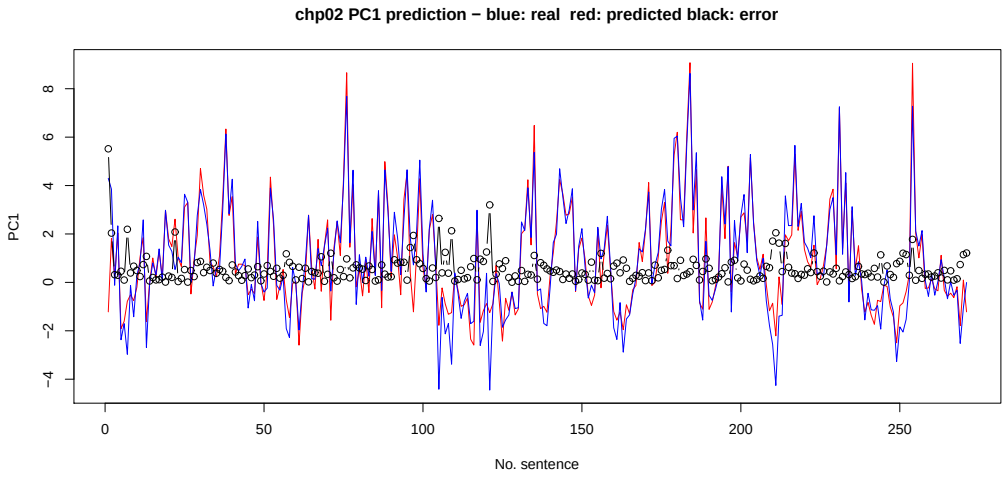

Figure 2: Prediction of PC1 using multiple linear regression of sentiment analysis scores and number of words in the sentence for chapter 02 of "The Adventures of Tom Sawyer".

ter 02 has 20.5 words in average per sentence (minimum 1 and maximum 93 words). These observations confirm that short sentences tend to be more expressive and therefore more unpredictable in terms of sentiment analysis (Mohammad, 2011). The sentences presented in Table 4, exemplify this difficulty, although from an acoustic point of view the model is able to capture quite well the style intended by the reader in the book. In fact auditively the sentences presented in this Table are quite different, which makes it possible to define and predict more than two expressive styles.

6. Conclusions

We have found that sentiment analysis scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean F0. Scores derived from lexicon and Sentiwordnet are much less correlated with the acoustic features in the analysed data. It is interesting to notice that any of the F0 contour measures (intonation measures) correlate with sentiment scores, this observation probably is in line with the findings of (Busso et al., 2009) where it has been found that gross pitch statistics are more emotionally prominent than features describing the pitch shape.

We have designed an experiment intended to predict the levels of acoustic expressivity in arbitrary text using sentiment analysis features and the number of words in the text. We have found that the predictive model fits well the training data, and it is able to predict the style of unseen data, in particular the character style of utterances in two chapters of the book not used for training the model.

An immediate application of these results is in automatic speech synthesis. We have demonstrated that an style can be automatically derived from textual data and a trained model, so the next step is to use this information to select the expressive style with wich the text should be realised. Also, given the clear auditive differentiation of utterances along PC1 values we will consider to predict more than two styles defining various PC1 thresholds.

7. Acknowledgements

This work is supported by the EU project SSPNet (FP7/2007-2013). We would like to thank Christopher Potts for providing us with the sentiment analysis of the data and Holmer Hemsen for assistance on data annotation.

8. References

Blizzard Challenge. 2012. Blizzard Challenge 2012. http://www.synsig.org/index.php/Blizzard_Challenge_2012.

Pang Bo, Lee Lillian, and Vaithyanathan Shivakumar. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA.

N. Braunschweiler, , and S. Buchholz. 2011a. Automatic sentence selection from speech corpora including diverse speech for improved hmm-tts synthesis qualityt=/. In *Interspeech*, Makuhari, Chiba, Japan.

N. Braunschweiler, M.J.F. Gales, and S. Buchholz. 2011b. Lightly supervised recognition for automatic alignment of large coherent speech recordings. In *Interspeech*, Makuhari, Chiba, Japan.

Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. *IEEE Transactions on Audio, Speech & Language Processing*, 17(4):582–596.

Wai C. Chu, 2003. *Mixed excitation linear prediction*, chapter 17, pages 454–485. Speech coding algorithms Foundations and Evolution of Standardized Coders. Wiley.

LibriVox. 2012. LibriVox acoustical liberation of books in the public domain. http://librivox.org/.

Bing Liu. 2011. Opinion mining, sentiment analysis, and opinion spam detection.

Character	Predicted_PC1	Text
narrator	3.00	Soon the free boys would come tripping along on all sorts of delicious expeditions, and they would make a world of fun of him for having to work -- the very thought of it burnt him like fire.
narrator	2.24	He then held a position at the gate for some time, daring the enemy to come outside, but the enemy only made faces at him through the window and declined.
narrator	1.69	If one moved, the other moved -- but only sidewise, in a circle; they kept face to face and eye to eye all the time.
narrator	0.58	So she lifted up her voice at an angle calculated for distance and shouted:
...		
narrator	0.05	Spare the rod and spile the child, as the Good Book says.
narrator	0.04	I reckon you're a kind of a singed cat, as the saying is -- better'n you look.
narrator	0.01	If you was to tackle this fence and anything was to happen to it -- "
other	-0.00	Another pause, and more eying and sidling around each other.
other	-0.00	Ben ranged up alongside of him.
other	-0.02	He opened his jacket.
other	-0.04	"Tom, it was middling warm in school, warn't it?"
other	-0.05	At this dark and hopeless moment an inspiration burst upon him!
...		
other	-1.87	"Nothing."
other	-1.96	"Aw -- take a walk!"
other	-1.97	I'll learn him!"
other	-1.98	"By jingo!
other	-2.11	"You can't."
other	-2.13	Course you would!"
other	-2.16	"Y-o-u-u TOM!"
other	-2.17	Oh, what a hat!"
other	-2.17	"Well why don't you?"
other	-2.18	Why don't you DO it?
other	-2.99	"Nothing!
other	-3.20	Ting-a-ling-ling!
other	-3.20	Chow-ow-ow!
other	-3.20	Ting-a-ling-ling!
other	-3.20	SH'T!

Table 4: Predicted PC1 value and corresponding text for some sentences of chapter 01.

http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html.

Saif Mohammad. 2011. From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA, June. Association for Computational Linguistics.

Christopher Potts and Florian Schwarz. 2010. Affective 'this'. *Linguistic Issues in Language Technology*, 3(5):1–30.

Christopher Potts. 2011a. On the negativity of negation. In Nan Li and David Lutz, editors, *Proceedings of Semantics and Linguistic Theory 20*, pages 636–659. CLC Publications, Ithaca, NY.

Christopher Potts. 2011b. Sentiment Symposium Tutorial: Lexicons. Section 3.4 Experience Project reaction distributions. http://sentiment.christopherpotts.net/lexicons.

Project Gutenberg. 2012. Free eBooks by Project Gutenberg. http://www.gutenberg.org/wiki/Main_Page.

R Development Core Team, 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

J. Reilly and L. Seibert. 2003. Language and emotion. In Richard J. Davidson, Klaus R. Scherer, and H. Hill Goldsmith, editors, *Handbook of Affective Sciences*, chapter 27, pages 535–559. Academic Press.

K. Sjölander. 2012. The snack sound toolkit. http://www.speech.kth.se/snack.

E. Székely, J. P. Cabral, P. Cahill, and J. Carson-Berndsen. 2011. Clustering expressive speech styles in audiobooks using glottal source paramters. In *Interspeech*, Florence, Italy.

L. Wang, Y. Zhao, M. Chu, Y. Chen, F. Soong, and Z. Cao. 2006. Exploring expressive speech space in an audiobook. In *Speech Prosody 2006*, Dresden, Germany.

Opinion and emotion in movies: a modular perspective to annotation

Mouka, E.[2], V. Giouli[1,2], A. Fotopoulou[1], I.E. Saridakis[2,3]

[1]Institute for Language and Speech Processing, R.C. "Athena", Greece;

[2]Ionian University, Greece; [3]National and Kapodistrian University of Athens, Greece;

f10mouk@dflti.ionio.gr, voula@ilsp.athena-innovation.gr, afotop@ilsp.athena-innovation.gr, iesaridakis@gmail.com

Abstract

This paper presents an ongoing effort work focusing on the development of an audiovisual corpus resource and its annotation in terms of sentiments and opinions. A modular annotation schema has been employed based on the specifications of existing schemas and extending or adapting them to cater for the peculiarities of the corpus-specific data.

Keywords: annotation of emotion, annotation of opinion, movies corpus

1. Introduction

This paper presents the first version of a new specialized audiovisual corpus resource that comprises movies coupled with both orthographic transcriptions in English [en] and their official subtitles in Greek [el] and Spanish [es]. The corpus resource bears annotations at various levels of analysis (word/phrase/sentence, and also on the audio) while the focus is on the identification of opinions and emotions in oral discourse, elaborating on specific semantic and pragmatic phenomena. Cross-language issues were considered as well as textual vs. audiovisual cues. We describe the specialized corpus focusing on the pilot annotation procedure, and the results of an inter-annotator agreement study.

The paper is organized as follows: Section 2 presents the scope and aims of the research undertaken, as well as the multimodal corpus in terms of its content and typology. Section 3 includes descriptions of the metadata. Section 4 presents an overview of related works. Section 5 includes a detailed overview of the annotation scheme employed and the methodology adopted in the current annotation work. Our preliminary findings are presented and discussed in Section 6. Finally, Section 7 includes our conclusions and prospects for future work.

2. Project scope and aims

The audiovisual data was initially selected in order to guide translation-oriented research examining the language options that depicts a specific type of biased opinionated ideological stance/attitudes, namely that of racist discourse, and its transfer from the textual source language (SL) to the target language(s) (TL) through subtitling. This type of discourse is socio-culturally marked (Waugh, 1982). This makes the corpus an excellent pool for annotating opinions, beliefs, thoughts, feelings, emotions, goals, evaluations, and judgments in oral discourse. In yet another aspect, the annotations were also oriented towards populating a lexical resource that is currently under development and contains opinion and emotion words with new entries adhering to oral data. The ultimate goal of this work, therefore, is to investigate the use of opinion and emotion expressions in oral discourse by means of a corpus annotation study that extends across modalities and languages.

Finally, this work which is still in progress may be integrated into a larger initiative undertaken by the Institute for Language and Speech Processing aimed at the development of a suite of language resources (corpora, lexica, tools) for sentiment analysis.

3. Corpus description

As a product of the so-called prefabricated orality (Baños-Piñero & Chaume, 2009), movies were selected according to external and internal criteria: (a) topic (centered around inter-racial relations); (b) time, i.e., contemporaneity of production and reference; (c) realistic approach to events; and (d) their content (assumed racist discourse). To date, the corpus comprises 5 movies with a total playtime of 09:05 hours of quasi-spontaneous oral speech. The [en] audio-visual material has been transcribed and segmented, and utterances have been synchronized (time aligned) with the movie audio. Finally, the transcripts were also aligned with "[el] and [es] subtitles from the official distribution of the movies. The subtitle material is a specialized type of translation corpus, in the sense that subtitling conforms to certain time and space restrictions.

More precisely, following standard procedures, so as to ensure conformity with standards for audio-visual material, and, thus re-usability of resources, video segmentation and transcription were performed using ELAN (Brugman & Russell, 2004). The spoken language transcription relied on the TEI specifications for (Schmidt, 2011; TEI Consortium, 2011). The segmentation was performed at the utterance level, following intonation and pause clues, while quite long stretches of speech were further segmented into C-Units to facilitate alignment with TL subtitles that follow written discourse conventions and typically comprise short sentences. Repetitions, hesitations, repairs and overlapping utterances that are inherent in oral discourse have been retained in the corpus. Each utterance is assigned a time slot and a speaker. The final output is a TEI-conformant .xml document. An example of the resulting representation is depicted in the Figure 1 below:

```
- <div>
- <u who="#SPK16">
  <anchor synch="#T835" />
     Who do you hate, Danny?
  <anchor synch="#T836" />
     </u>
- <spanGrp type="subtitles" lang el">
  <span from="#T835" to="#T836"> Ποιον μισείς,
     Ντάνι ;</span>
     </spanGrp>
- <spanGrp type="subtitles es">
    <span from="#T835" to="#T836">¿A quién
     odias, Danny?</span>
     </spanGrp>
     </div>
- <div>
```

Figure 1: transcribed text

The external structural annotation (including text classification) of the corpus also adheres to the IMDI metadata scheme (IMDI Team, 2003). IMDI metadata elements for catalogue descriptions (IMDI Team, 2009) were also taken into account to render the corpus, and adaptations proposed specifically concerning Multimodal Language Resources have been taken into account. This type of metadata descriptions was added via the ELAN interface and stored in XML format.

4. Opinion and emotion: background

Background work has for the most part focused on sentiment classification, at the document, sentence or even phrase and word level. The MPQA corpus of news documentation (Wiebe et al., 2005; Wilson, 2008) defines attitudes as private states and proposes an annotation schema catering for the following conceptualizations or types of attitude: *sentiment*, *agreement*, *arguing*, *intension*, and *speculation*. A general type, posited as *other attitude* is retained for all the remaining private states and a value of *positive* or *negative* is also assigned to the specific classes, as well as fine-grained intensity values. Expressive subjective elements, subjective speech events and explicit mention of private states are annotated separately. Agents and targets are also considered. Somasundaran & Wiebe (2010) explore further the *arguing* type as a means to investigate ideological stance. Opinion-target pairs are created, encoding also what the opinion is about, on the basis that opinions combined with targets are more informative than either of them in isolation.

Asher at al. (2009) have worked on a corpus compiled by movie reviews, letters to the editor and news reports to define a fine-grained annotation scheme that builds on the semantics of a wide class of opinion expressions at the sub-sentential level, the latter ultimately mapped onto a top-level typology of *reporting* (indicated by verbs), *judgement* (that builds on the semantics of a wide class of opinion expressions at the sub-sentential level), *advise* and *sentiment* expressions. This scheme is argued to be appropriate for calculating the overall opinion expressed in a text on a given topic.

The *Emotiblog* annotation model has been used on a corpus of various textual genres (news articles, news titles and a corpus of real-life self-expressed emotion) (Boldrini et al., 2010) and a corpus of blogs (Balahur et al., 2010) and distinguishes between objective and subjective speech. Polarity is assigned to adjectives/adverbs, verbs, nouns, anaphora and orthographic features. Interestingly, it takes into account two attributes (reader and author interpretation), annotating cases where apparently objective statements are used as indirect expressions of opinion.

As far as polarity is concerned, Polanyi & Zaenen (2006) examine how lexical valence is context-dependent and how valence shifters, such as negatives/intensifiers, modals, irony and various discourse structures influence the polarity and/or the strength of the opinion expressed. Furthermore, Neviarouskaya et al. (2010), based on the *Appraisal Theory* (Martin & White, 2005) present a scheme that includes polarity (positive, negative, neutral) on the top level, which is further divided into three types (affect, judgment and appreciation). Affect is further subdivided into 8 basic types. The authors propose an algorithm to decide how polarity is affected by a set of attitude-conveying terms, modifiers, functional words and modal operators. Using the *compositionality principle* the overall meaning of a sentence is determined.

Finally, the Boloscopy corpus (Daille et al., 2011), containing personal thematic blogs, is annotated according to five types of evaluations: *opinion* (conviction/supposition), *appreciation*, *acceptance -refusal*, *agreement-discord* and *judgement*. Implicit and explicit cases are taken into account, as long as positive/negative polarity

5. Opinion Annotation in Movies

In this section we will elaborate further on the annotation schema employed that caters to the identification of two broad categories: (a) *emotion*, expressing the psychological state of a speaker or an agent towards somebody or something usually based on feeling or sentiment rather than reasoning; and (b) *opinion*, that is an expression of attitude, speculation, beliefs, thoughts, etc. The schema, therefore, comprises two basic elements, namely, *emotion* and *opinion*.

The schema also considers a more fine-grained classification of opinion and sentiment. *Emotion* classification is centred around a set of 8 basic sentiments (Plutchik, 1991): *anger*, *fear*, *sadness*, *disgust*, *surprise*, *anticipation*, *acceptance*, *joy* and *other*. Moreover, the following opinion classes are defined: *evaluation*, *belief*, *recommendation*, *intention* and *other*. More precisely, an evaluation is specified as an estimation of the value of a person, object, action, etc., an assessment of behaviour or of phenomena, and involves both ethic and aesthetic values. Under the umbrella term *belief* we classify expressions denoting the point of view of the speaker, of what he believes to be true, possibly used as an argument. Additionally, *intentions* encompass aims, plans and other

overt expressions of intention, while *recommendations* are further defined as expressions intenting to urge the interlocutor to take an action.

Polarity of sentiment/opinion was also assigned to the selected text spans (being either sentences/clauses or phrases/words) as a mandatory feature assuming one of the following values: *positive*, *negative*, *neutral*, and *uncertain*.

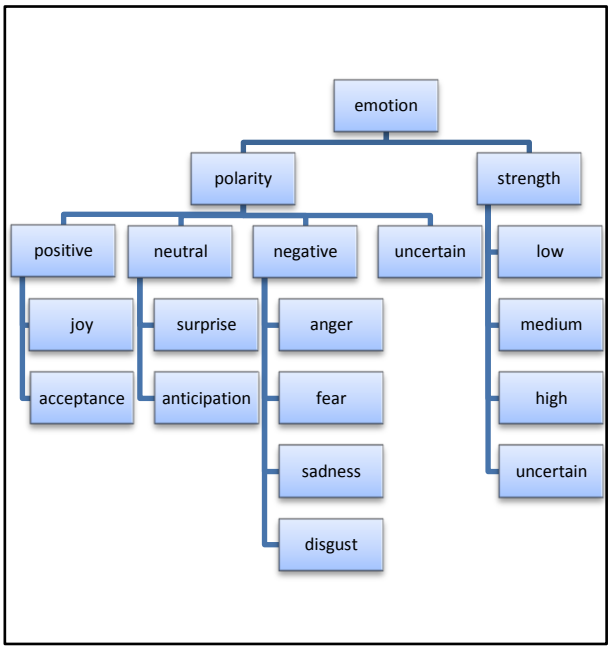

Figure 1: Emotion annotation schema

Emotions and opinions were further assigned the mandatory feature of *strength* the possible values of which are: *low, medium, high.* An extra value *uncertain* was also provided for, in order to make annotators assign a value only if they are sure, leaving difficult or

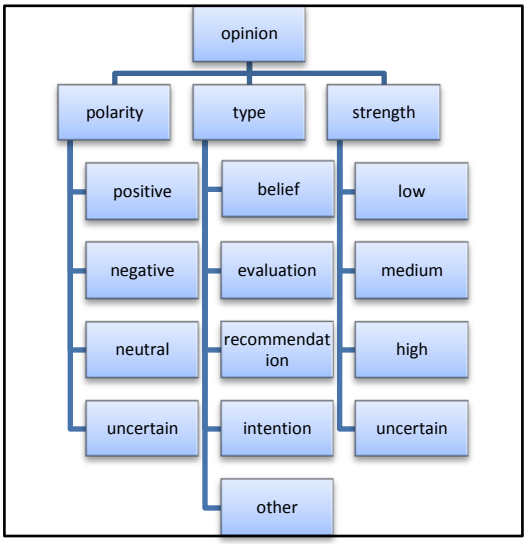

Figure 2: Opinion annotation Schema

ambiguous cases for future treatment. Lexical choices and/or paralinguistic features (especially for emotions) denote the strength of the opinion or emotion and are therefore taken into account during the annotation process. The annotation schema for elements emotion and opinion depicting mandatory features and their possible values is depicted schematically in Figure 1 and Figure 2 respectively.

Moreover, three more features that are optional are proposed in the schema, namely *inferred*, *repetition* and *irony*. The feature *inferred* has been employed to distinguish utterances for which the opinion or emotion values are assigned on the basis of non-verbal evidence or paralinguistic cues. Possible values for this feature are: *audio*, *video*, *text*, that is, the modality contributing to the interpretation of the utterance. Additionally, the feature of *repetition* has also been used to mark cases where the repetition of an utterance or part of it (on word, syntax or phrase level) is used to express the strength of an emotion or opinion (that is *set to yes if true, otherwise it is assigned the value no).* The feature *irony* is used to encode pragmatic phenomena (see section 5.1).

The proposed schema builds on existing annotations of emotion and opinion in discourse with certain modifications that were deemed necessary so as to accommodate the peculiarities of the data at hand, namely phenomena that are inherent to oral discourse.

Speakers express their own or other persons' emotional states, opinions, evaluations, etc, either explicitly or implicitly. For example, in utterances (1) - (3) below, the speaker's emotional state is expressed directly, using an emotion expression:

(1) Danny, no, <emotion>I feel sorry for you, Danny </emotion>.
(2) <emotion> I hate anyone that isn't white Protestant</emotion>.
(3) <emotion> I am angry</emotion> all the time.

In (4) the speaker explicitly expresses his belief or speculation, the correct interpretation of which is based on the modal *wouldn't*, whereas in (5) the verb "*think*" further reinforces the interpretation of the utterance as an opinionated one:

(4) <opinion – speculation> Derek wouldn't let us visit him in prison </opinion – speculation>
(5) <opinion – speculation> I think the street would kill you</opinion – speculation>

However, emotional states or opinionated discourse may be expressed implicitly as well. Interlocutors usually make use of implicit lexical choices to express their attitudes, as for example in utterances (6), or they make use of paralinguistic cues to express their emotional states, as in (7):

(6) <emotion-joy>
Good to have you back </emotion-joy>.
(7) Danny! Danny! <emotion-anger> Danny</emotion–anger>!
<emotion–anger>Shut the door</emotion–anger>!

To accommodate these cases, a further feature *inferred*

with possible values "*audio", "video", "context"* has also been added in order to retain information on the modality that contributes to the correct interpretation of the utterance (see above). The identification of implicitly expressed emotions is not a trivial task. In fact, this is even harder in cases of overt opinions which convey also the speaker's emotional load. In this work, however, we have attempted to annotate utterances which are opinionated and at the same time express the emotional state of the speaker:

(8) This is typical. (opinion-evaluation, emotion-sadness)

(9) This country is becoming a heaven for criminals. (opinion-belief, emotion-anger)

(10) This isn't fair! (opinion-evaluation, emotion-anger)

(11) They are the fucking enemy. (opinion-evaluation, emotion-disgust)

Finally, the annotation specifications allow the annotation of nested opinions and emotions, as shown in the following examples:

(12) <opinion-belief>
<opinion-evaluation>Decent, hardworking Americans, like my dad </opinion-evaluation> are rubbed out by <opinion-evaluation> social parasites</opinion-evaluation>
</opinion-belief>.

(13) But <opinion-belief> if a white person sees two black men walking towards her, and she turns and walks in the other direction, <opinion-evaluation> she's a racist </opinion-evaluation> </opinion-belief>, right?

(14) I know <opinion-belief> you don't believe any of <opinion-evaluation> this shit </opinion-evaluation></opinion-belief>, right?

5.1 Annotating pragmatic phenomena

Movies comprising our corpus depict situations in which dialogue participants make use of a wide range of communicative and rhetorical devices. To render the subjectivity annotation as complete as possible, pragmatic phenomena were also taken into account and irony was the first one to be annotated. Ironic/sarcastic utterances were also identified and marked as appropriate.

Irony is generally defined as a form of *non-sincere speech*, as a means to convey a meaning which is *opposite* or *different to the literal one*, and has been treated as a violation of the Gricean Maxims, principally of that of Quality (Alba Juez, 1995). According to the Maxim of Relevance, listeners attempt to interpret non-explicitly relevant utterances in a manner that fulfils the expectation of relevance and are thus able to recognize the ironic dimension in speech. From another perspective, irony has been proven to function in both a positive and a negative way. In Alba Juez (1995) two main kinds of irony were proposed: *Positive Irony* (intended to praise) and *Negative Irony* (intended to criticize). The annotation scheme that we have developed, takes this double classification into account, however, only one instance of positive irony has been identified so far, and marked appropriately as "irony-positive", in (15) where irony is used to imply that the boy is an excellent student. Examples in (16), (17), and (18) are all cases of negative irony/sarcasm which clearly show how the literal lexical meaning is altered by irony

(15) What's the matter, afraid you're going to get a B?

(16) Give yourself a raise, will you? (while depreciating the work of his colleague)

(17) Hey, that's a *great* color on you, you know? Now you can get a white woman's job, bitch.

(18) This country is becoming a *heaven* for criminals.

Annotation of irony in the corpus was performed on the basis of contextual and/or world and situation-specific knowledge. Moreover, since our data involves the oral modality, identification of ironic utterances was also aided by acoustic features. On the basis of the assumption that speakers provide prosodic disambiguation cues when using verbal irony and that listeners use prosodic information, in addition to context information, to interpret ironic utterances (Bryant & Fox Tree, 2002), intonation was also used as a cue for disambiguation.

5.2 Annotation methodology

After the initial specifications were formulated, annotations as outlined above were applied by three expert linguists separately for each language and modality in a *modular* way. More precisely, annotation was initially performed on the [en] transcripts at the phrase and word level, first assigning a polarity. Nouns, adjectives, adverbs, verbs and multi-word expressions were treated. Further annotation was then performed at the sentence and clause level. We did not provide annotators with any predefined grammatical categories and the span of every annotation corresponds to the extended units of meaning (Sinclair, 1996; Hunston, 2007), i.e. what fragment of text is considered to express each emotion or opinion in the communicative instance

At the next level, cues beyond lexis that were provided by the audiovisual material were also taken into account with respect to the speakers' emotional state. To this end, a second round of annotation was initiated with annotators taking into account acoustic and visual cues, such as intonation, gestures and body language to interpret utterances.

As noted already, this procedure has been conceived of as a modular approach to annotation. Each level (word/phrase, clause/sentence) or modality contributes separately to the overall emotional load or attitude expressed either in a film or in any given film scene, shot, etc. To keep track at any given point of the contributing level or modality, however, each text span has been coupled with information on the level or modality from which the opinion or emotion is inferred. Many applications would benefit from being able to determine not just whether a film or scene is opinionated or emotionally overloaded, but also the contributing level or modality.

The source of every speech event is by default the

corresponding speaker and therefore it is not explicitly identified. Although we consider identifying targets, annotation at this level has not been implemented yet.
Finally, annotation was performed using the GATE (version 7.0) platform (Cunningham et al., 2002). The tool was selected for its user-friendliness and its versatility in fulfilling all the requirements of our annotation model.

6. Discussion

For the time being, only two movies have been annotated in two of the languages involved in the study: en-transcripts along with their el-subtitles.
It should be noted that, as one might expect, the data included in our corpus is quite different from the data usually treated in similar efforts that have been reported in the literature (see section 4), in that our data is oral and includes a significant amount of implicit speech events (not triggered by expressions, such as "I said", as in other works), conversational and includes highly colloquial discourse elements. This is a unique feature of our textual evidence.
To ensure annotation quality in terms of consistency, and in view of identifying problematic cases, inter-annotator agreement was calculated using Cohen's kappa coefficient (Cohen, 1960), i.e. a statistical measure of inter-rater measure for qualitative items. In an evaluation experiment involving 50 utterances, the inter-annotator agreement between 2 separate annotators on the word/phrase level was 0.92, dropping significantly at the sentence level (0.67 when all features were considered, and 0.86 when only polarity was taken into account).
Admittedly, annotating opinion and emotion in text is not a trivial task. Agreement was achieved in clear-cut cases, as in the following examples:

(19) Sweeney's *a good teacher*. (opinion – positive)
(20) I'm telling you, man, this kid is *smart*. (opinion – positive)
(21) this kid is *a genious*. (opinion – positive)
(22) Sweeney is *a nigger on a power trip*, (opinion – negative) Vinyard.
(23) They're *a burden* to the advancement of the white race. (opinion – negative)
(24) The gangs are *like a plague*. (opinion – negative)
(25) You've got to draw the line. (recommendation-neutral)
(26) I'm not ready to give up on him yet. (intention positive)
(27) I can guarantee you one hundred per cent his brother did not put him up to this. (opinion-belief, polarity-positive, strength-high)
(28) You hate this child (emotion – negative)

Instances presenting a disagreement between annotators must be further analysed so as to explicate the reasons underlying this difference. However, in cases which seem to be the most problematic, sentiment is not directly concluded from the co-text of the utterance examined.

7. Conclusions and Future Work

This paper has presented our ongoing work, aimed to develop a new specialized multimodal resource and to implement a pilot annotation scheme, so as to identify and represent opinions and emotions from a multi-modal perspective. The resource will tentatively be useful in a variety of cross-language studies and applications
Future work involves annotating of the remaining video material, including the [es] subtitles and developing and implementing a more fine-grained annotation scheme for our audio and video material, especially with respect to pragmatic phenomena, in order to facilitate a comparison between source and target texts and draw conclusions on translational norms and behaviours (Toury, 1995; Saridakis, 2010) with regard to subtitling practices in Greece and Spain.
Moreover, following common practices (Wiebe et al., 2005; Wilson, 2008), additional features will be implemented, as for example the identification of opinion/sentiment frames that consist of the opinion-holder or sentiment-experiencer and the target of opinion/sentiment respectively, etc.
Our future plans include also the annotation of the textual material at the various levels of linguistic analysis, with the focus being on the syntax and semantics of verbs, nouns, and adjectives that are indicative of emotions and/or opinions.
In conclusion, the present work might also prove useful for other researchers interested in the multimodal annotation, in the fields of sentiment and subjectivity analysis.

8. Acknowledgements

We must thank the anonymous reviewers for their useful suggestions and comments. Part of this work was supported by the Greek State Scholarships Foundation (IKY) through a Ph.D. scholarship awarded to E. Mouka.

9. References

Alba Juez, L. (1995). Irony and Politeness. *Revista Española de Lingüística Aplicada*, 10, pp 9-16.
Asher, N., Benamara, F., Mathieu, Y.Y (2009). Appraisal of opinion expressions in discourse. *Lingvisticae Investigationes*, 32(2).
Balahur, A., Boldrini, E., Montoyo, A., Martinez-Barco. P. (2010). IBEREVAL OM: Mining Opinions from the new textual genres. *Procesamiento del Languaje*, 45, pp. 267-271.
Baños-Piñero, R. & Chaume, F. (2009). Prefabricated orality. A challenge in audiovisual translation. *Intralinea*, Special Issue: The translation of dialects in multimedia.
Boldrini, E., Balahur, A., Martínez-Barco, P., Montoyo, A. (2010). EmotiBlog: a finer-grained and more precise learning of subjectivity expression models. In *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV), satellite workshop to ACL 2010*. Uppsala

University.

Brugman, H., A. Russell (2004). Annotating Multimedia / Multi-modal resources with ELAN . In: *Proceedings of LREC 2004, 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon.

Bryant, G.A., & Fox Tree, J.E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor and Symbol*, 17(2), pp. 99-117.

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement*, 20 (1), pp.37-46.

Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V. (2002). GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).* Philadelphia.

Daille, B., Dubreil, E., Monceaux, L., Vernier, M. (2011). Annotating opinion - evaluation of blogs: the Blogoscopy corpus. *Language Resources and Evaluation*, 45(4), pp. 409-437.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics.* 12.2, pp. 249–268.

IMDI Team, (2009). IMDI Metadata Elements for Catalogue Descriptions, Version 3.0.13, MPI Nijmegen.

IMDI Team. (2003). IMDI Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen.

Martin, J.R. & White, P.R.R. (2005). The Language of Evaluation: Appraisal in English. Palgrave, London, UK.

Neviarouskaya, A., Prendinger, H., Ishizuka, M. (2010). Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).* Coling 2010 Organizing Committee, Beijing, China, pp. 806-814 .

Plutchik, R. (1991). *The Emotions*. New York, NY: University Press of America.

Polanyi, L., Zaenen, A. (2006). Contextual Valence Shifters. In J. Shanahan, Y. Qu, and J. Wiebe (eds.) *Computing Attitude and Affect in Text: Theory and applications.* The Information Retrieval Series, Vol. 20, Springer, Dordrecht, The Netherlands, pp.1-9.

Saridakis, I.E. (2010). *Text Corpora and Translation. Theories and Applications*. Athens: Papazisi (in Greek).

Schmidt, T. (2011). A TEI-based Approach to Standardising Spoken Language, *Journal of the Text Encoding Initiative*, 1.

Sinclair, J. (1996). The search for units of meaning. *Textus*, 9.1, pp. 75–106.

Somasundaran, S. & Wiebe, J. (2010). Recognizing Stances in Ideological On-Line Debates. In *NAACL-HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text.* Los Angeles.

TEI Consortium, (eds.) TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.3.0. May 2011. TEI Consortium.

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Waugh, L. (1982). Marked and Unmarked: A Choice Between Unequals in Semiotic Structure. *Semiotica* ,38, pp. 299-318.

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

Wilson, T. A. (2008). Fine-grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States. Ph.D. thesis, University of Pittsburgh.