# SESSION P18-S

## Speech Corpora & Annotation/Processing Tools

# Reusing Language Resources for Speech Applications involving Emotion

**Christina Alexandris and Stavroula-Evita Fotinea**

Institute for Language and Speech Processing
Artemidos 6 & Epidavrou, GR 151 25, Maroussi, Greece
{calex,evita}@ilsp.gr

**Abstract**

The present paper involves using a spoken corpus for the construction of a written corpus which in turn will be used for speech applications involving emotion, namely an emotional Text-to-Speech system or a Speech-to-Emotion system which requires emotional speech recognition and consequent text to emotion conversion. Such speech application systems, involve the construction of a corpus of written artificial dialogs, which is intended to depict/convey the speakers emotional state. From the analysis of the sublanguage of the spoken corpus as well as the analysis and evaluation of the sublanguage of the (constructed) written corpus, we classified the extra-linguistic markers into three types ('Pause', 'Emphasis' and 'Hesitation') and we observe that, in Greek, extra-linguistic markers may behave as pointers to key-information. The correspondence between extra-linguistic markers and key-information was evaluated with an additional spoken corpus of recorded dialogs from selected Greek TV programs. The written corpus containing the inserted extra-linguistic markers was evaluated by native speakers with linguistic knowledge. The speakers were asked to classify them according to their acceptability in Greek (Alexandris & Fotinea, 2003).

## Introduction

Emotion presents a challenge in both Speech Recognition and Speech Synthesis systems since it constitutes an element that is heavily speaker dependent. However, the definition of emotional states may be more accurate if the element of emotion is subject to a set of constraints. The present paper involves the constraints of sublanguage and speech-act as well as the constraint related to speaker type.

For the simulation of emotions in a Text-to-Speech system we use a spoken corpus in the domain of a specific sublanguage as a basis for detecting and extracting emotional features. We adapted the extracted emotional features to the framework and sublanguage of the TTS system, a task that involves the construction of a written corpus. In the written corpus we use extra-linguistic markers as prosodic markers related to key-information (Alexandris, 2003), for the depiction of the speaker's emotional state.

The analysis of the behaviour of prosodic markers in the Greek language was based on the study of a spoken corpus of journalistic texts (CIMWOS Project) recorded from television and subsequently transcribed and annotated according to a given set of standards used for the spoken corpora of the specific project. The spoken corpus was used for the extraction of extra-linguistic markers for their subsequent insertion in the appropriate position in the written corpus. The extra-linguistic markers were inserted in positions related to 'keywords' in the written corpus of the TTS system. The categorization of the extra-linguistic markers and their mapping to the respective type of emotion was based on our analysis of the spoken corpus.

## Analysis and Preprocessing of the Speech Corpus

### Outline of Tasks

Two tasks were performed on the transcribed and annotated spoken corpus of the CIMWOS Project (2000-2003), namely (1) the analysis of the sublanguage in respect to discourse structure and speech acts and (2) the categorization of speakers.

### Analysis of Sublanguage

#### Sublanguage and Discourse Structure

The sublanguage of the present spoken corpus (Figure 1), namely its lexical, syntactic and semantic characteristics (Hoffmann , 1989 ; Kelz, 1983) is not a sublanguage limited to a small group of word classes and syntactic structures. The core of the semantic content of the sentence may be expressed in the form of questions i.e. "Who (Person)" or "What(Action)". The type of questions or relations expressed depend on the speech act performed. Therefore, the core of the semantic content of every sentence is defined in terms of criteria related to the discourse structure of its context.

ERT_20020510_1958_Net.trs, CIMWOS Project
SPEAKER[Minister Georgios Floridis] :
What we decided about recruitments is that [PAUSE] they will obey to two rules, first {of all} to the necessity [BREATH] of an [EH] effective operation of the [AH] government. And to offer better services to the citizens. And, second[Mispronounced] {of all}, that {all} these recruitements [PAUSE] should be [EH] according to the possibilities of the budget {in question}.

Figure 1: Speech Corpus example (Translation parallel to the Greek Text).

#### Definition of Speech Acts

The basic speech acts in the corpus are divided into four types, namely 'Announcement', 'Description', 'Declaration' and 'Expression'. The 'Announcement' speech act involves the announcement of news. This speech act is performed by journalists (anchormen and correspondents) only. The 'Description' speech act involves the description of a situation. This speech act is performed by all speakers except politicians. The 'Declaration' speech act is related to the declaration of one's opinion, position or the announcement of facts. The

'Declaration' speech act is performed by politicians and by non-politicians, namely professionals such as government officials, police chiefs, doctors, experts in a scientific field. The 'Expression' speech act involves the expression of one's feelings and is limited to speakers that are neither journalists nor politicans.

**Formal Analysis of Speech Act Topic & Content**

The discourse analysis of the corpus was performed manually. However, we used templates as a formal framework for the analysis and presentation of the topic and content of each speech act in the discourse framework of the corpus. Each template corresponds to a specific speech act ("SPEECH-ACT") which has a particular topic ("TOPIC") or a set of topics. Each topic has a specific content ("WHAT") which may be one or more items ("WHAT-1", "WHAT-2" etc.).

In the following example (Figure 2), the speaker, a politician, performs a 'Declaration' speech act in declaring the following topic (signalized as "TOPIC" in the discourse analysis performed), namely the decisions of the Ministry of National Economy in respect to staff recruitments. The content of the topic of the declaration (signalized as what "WHAT") involves "two rules", analysed as "WHAT-1" and "WHAT-2".

ERT_20020510_1958_Net.trs, CIMWOS Project
SPEAKER[Minister Georgios Floridis] :
SPEECH-ACT: DECLARE
TOPIC: What we decided about recruitments is that
[PAUSE]
WHAT
they will obey to two rules, first {of all} to the necessity [BREATH] of an
WHAT-1
[EH] effective operation of the [AH] government. And to offer better services to the citizens. And, second[Mispronounced] {of all}, that {all} these recruitments
[PAUSE]
WHAT-2
should be [EH] according to the possibilities of the budget {in question}.

Figure 2: Analysis in respect to the Speech acts.

## Speaker Categorization

The speakers of the transcribed and annotated spoken corpus are divided in two major categories, Trained Speakers (I) and Non-Trained Speakers (II). The Trained Speakers category consists of three groups, namely Journalists, Politicians and Actors. The Journalist group, which constitutes the largest group of trained speakers, is divided into the 'Anchormen' and 'Correspondents' subgroups. The Actors group is the smallest group in the data. The Non-Trained Speakers category consists of the subgroups Category 1 & 2. Category 1 involves speakers who are not trained but clarity and consistency in their speech is required by their profession or by circumstances related to their profession. This subgroup consists mainly of officials in the public sector and scientists. Category 2 involves all other non-trained speakers.

## Classifying Extra-Linguistic Markers in the Spoken Corpus

We divided the extra-linguistic markers into three categories, according to their corresponding type of speech signal. The first category of extra-linguistic markers (Type 1) consists of a pause before the word or phrases uttered by the speaker. The pause the speaker makes is transcribed as SKIP(PAUSE) or SKIP(OTHER), if there is background noise. The second category (Type 2) involves the emphasis of the word or phrase where, among other prosodic characteristics, an increase of the volume intensity of the speaker's voice is recorded. The third category of extra-linguistic markers (Type 3) consists of the sounds "Ah", "Eh" and "Ih", transcribed as [AH], [EH] and [IH]. The "Ih" sound is typical of Modern Greek.

The core of the semantic content of the sentence, defined by the previously described sublanguage of the spoken corpus is often emphasized by the speaker (Type 2) or preceded by a pause (Type 1). An equally frequent phenomenon in the present spoken corpus are the sounds [AH], [EH] and [IH] (Type 3) preceding the core of the semantic content of the sentence. In spoken Greek, the sounds "Ah", "Eh" and "Ih" are related to the hesitation of the speaker. We, therefore, name the third category of extra-linguistic markers the 'Hesitation' category.

## Type of Extra-Linguistic Markers in Respect to Speech Act and Speaker

From the analysis of the spoken corpus we observe that the type of extra-linguistic marker varies according to the speaker category and the speech act performed. Specifically, we observe that all speakers used Hesitation-markers before key-information when performing a speech-act of Description and Hesitation-markers and /or a pause before key-information when performing a speech-act of Declaration and Expression. Non-trained Speakers of Category 2 also used Emphasis to express their feelings. Emphasis was the only type of extra-linguistic marker used before key-information in Announcement speech-acts, exclusively performed by the Journalist Trained Speaker group (Figure 3).

| Speech Act<br>Speakers | An-nounce | Descri-be | Declare | Express |
|---|---|---|---|---|
| Journalist Anchor | E | - | - | - |
| Journalist Correspondent | E | H | - | - |
| Politician | - | - | P<br>H | - |
| Non-trained Speakers Category 1 | - | H | P<br>H | H<br>P |
| Non-trained Speakers Category 2 | - | H | - | H<br>E |

Figure 3: Distribution of Type of Extra-Linguistic Markers (E: Emphasis, H: Hesitation, P: Pause) in Respect to Speech Act and Speaker Category.

The above described correspondence was also confirmed by an additional corpus of recorded dialogs from Greek TV, namely a set of Journalistic talk-shows, especially recorded and analysed for that purpose. This correspondence can also be suppored by lexical and discourse elements (i.e. politeness) (Alexandris & Fotinea, 2003) depicting the emotional tone of the sentence.

## Construction of a Text Corpus Using Elements of the Speech Corpus

### Insertion of the Extra-Linguistic Markers in the Text Corpus

The reconstruction process involves the semi-automatic insertion of the extra-linguistic markers extracted from the spoken corpus in the appropriate positions in the written corpus of constructed artificial dialogs. The extra linguistic markers function as prosodic markers in the constructed corpus for Speech Synthesis applications. The entire system works in the framework of a sublanguage for on-line weather-reports. The system has "feelings" and expresses them to the user. When the weather is bad, the system is skeptical/melancholic and when the weather is good, the system is happy. The creation of feelings in the system are used as a basic framework for a future further development involving the detection of the users emotions and its adaptation to them.

The extra-linguistic markers are inserted in the appropriate positions in the sublanguage, namely in respect to the keyword. Keywords constitute a basic element in speech processing systems (Abela & Gorin, 1997). Specifically, they can be used in the language-processing module (i.e. prosodic-syntactic grouping), which consists a necessary module in state-of-the-art TTS systems, in addition to acoustic-linguistic processing and the final digital signal processing stage (speech synthesis) (Dutoit et al., 2000). Most Text-to-Speech systems include prosodic instructions computed by the TTS system for encoding appropriate linguistic and paralinguistic information (Bailly et al., 2000).

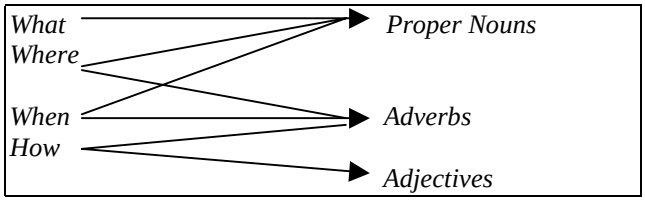


Figure 4: Basic Types of Information of the Text Corpus Sublanguage

The basic types of information contained in the sentences of the sublanguage of the text corpus may be expressed in the form of questions i.e. "What(Weather-Phenomenon)", "Where(Place)", "When(Time)" and "How(Degree)" (Figure 4). A specific type of information is related to a specific class of keywords. For instance, the keyword "sea" is classified as a "Weather-Phenomenon-Location" type (other examples are the keywords "mountain" and "plain" as particular parameters of the category "Where(Place)"). The "Weather-Phenomenon-Location" type "sea" is related to the "How(Degree)" keyword types with possible values "calm", "few waves", "slightly turbulent", "stormy" and "very stormy" (Figure 5).

Where: Place = Samos
What: Weather-Phenomenon = rain
How (Degree = heavy)
Weather-Phenomenon-Measure= windspeed
How (Degree = 8 Beaufort)
Weather-Phenomenon-Measure= temperature
Weather-Phenomenon-Location = sea
How (Degree = stormy)

Figure 5: Examples of Keywords in the Text Corpus Sublanguage

The extra-linguistic markers of 'Hesitation' and 'Pause' are inserted before the keywords and the extra-linguistic markers of 'Emphasis' are used on the keywords.

The extra-linguistic markers of the 'Hesitation' category are used in emotional dialogs expressing skeptical mood/melancholy. The extra-linguistic markers of the 'Emphasis' category are used in the dialogs expressing happiness (Scherer, 2000) (Figure 6, Figure 7). We note here that the intensity of 'Emphasis' in the emotion of happiness may be language-specific for Greek.

| *Emotion* | *Inserted-Marker* |
|---|---|
| Happiness | Emphasis |
| Skeptical | Hesitation |

Figure 6: Distribution of Type of Extra-Linguistic Markers in Respect to Emotional State.

[ SYSTEM- SKEPTICAL ]
SPEAKER: Tell me about <Samos>.
SYSTEM: In Samos there is heavy rain,
the temperatures have [EH] dropped, the windspeed is [AH] at 8 Beaufort and the sea is [EH] stormy.

[ SYSTEM-HAPPY ]
SPEAKER: How is the weather like in <Rhodes>?
SYSTEM: In Rhodes the temperatures is high, the windspeed is 2 Beaufort and the sea is very calm.

Figure 7: Example of Constructed Dialog.

### Evaluation of the Text Corpus and Results

The constructed dialogs were subsequently read aloud by one male and one female speaker and recorded. Thus, the written corpus was both in text and audio (wave signal) form. The recorded (80) dialogs were evaluated by native speakers with linguistic knowledge.

The native speakers were asked to classify the dialogs they were given in respect to three emotional states, namely "happiness", "skeptical /melancholy" and "neutral". The questionnaire was given in the form of multiple-choice questions. The native speakers were given four variations of each dialog. The variations were in random order. One dialog did not contain the extra-

linguistic prosodic markers and three variations of one dialog containing the extra-linguistic prosodic markers. One variation was heavily marked with the extra-linguistic prosodic markers, in another variation the marking was discrete and in a third variation the marking was of intermediate intensity. The Greek native speakers that evaluated the "naturalness" of the dialogs were asked to mark each dialog as "naturally-sounding" or "not naturally-sounding". 89% of the speakers perceived the dialogs with extra-linguistic markers as "naturally-sounding". The preferred variation of the dialog containing extra-linguistic markers and its respective degree of naturalness was not evaluated here.

The present categorization of extra-linguistic markers serves as a general guideline to outline tendencies in respect to expressing emotional state in Modern Greek. It must be noted that the perception and classification of emotion from the evaluators is dependent on factors such as the personality, age and sex of the individual. We note here that women tend to demonstrate a stronger sensitivity in respect to the expression and perception of emotion (Wardhaugh, 1992).

## Conclusion and Further research

From the behaviour of the extra-linguistic markers described in the present analysis we conclude that in Greek, extra-linguistic markers behave as pointers to key-information in two sublanguages, namely, in the broad field of journalistic texts and in the restricted sublanguage of weather reports.

The present categorization of extra-linguistic markers may be used in a preprocessing module for Greek that can operate within a Greek Text-to-Speech Synthesis system. The texts are enriched with prosodic markers before their processing by the TTS. The preprocessing module can contribute to the production of more naturally sounding synthetic speech with an easily intelligible content. We note here that in the planned future development of the system, emotions such as anger and fear will be incorporated.

DIALOG 5, ENGLISH & GREEK DIALOGS,
VERSION: 15/09/2003, ERMIS PROJECT, ILSP

SYSTEM: <Come on, you surely have something to say.>
SPEAKER: [ER] same old stuff.
SYSTEM: <Do you lead a stressful life?>
SPEAKER: It's not the most [EH] stressful life imaginable but PAUSE [well yes] it does have its [ER] stressful moments.

Figure 8: Example of a dialog of a system under development (ERMIS Corpus).

Extra-linguistic markers may be used in User Emotion Detection systems (Figure 8), in Information Retrieval systems involving spoken language etc. They may be used in dialog systems in services where the detection of emotions is important such as hospital services and applications for differently-abled people. Furthermore, extra-linguistic markers may serve as diagnostic tools for discerning the intentions of the speaker (or writer) and outlining a general (or specific) identity profile for speaker identification and forensic purposes.

Further research from a cross-linguistic aspect includes determining whether extra-linguistic markers behave as pointers to key-information in other languages and, if so, whether these markers demonstrate similarities and/or differences in respect to the types of extra-linguistic markers (for Greek) presented in the present paper.

**Acknowledgements**: The authors wish to acknowledge the assistance of Mr. Nikos Maganiotis, Synchresis Studios Ltd., Lakedaimonos Str. 9, GR 11523, Athens, Greece.

## References

Abela, A. & Gorin, A. L. (1997). Generating Semantically Consistent Inputs to A Dialog Manager. In Proceedings of Eurospeech 1997 (pp.1879--1882), Rhodes, Greece.

Alexandris, C. (2003). Using Sublanguage as a Tool for Determining Focus Elements in Greek Spoken Journalistic Texts on International News, in: Studies of the Greek Language, Thessaloniki 9-11 May 2003 (in Greek, in print).

Alexandris, C. & Fotinea, S.-E. (2003). Using Discourse Particles as Indicators of Positive Politeness in the Discourse Structure of Dialog Systems for Modern Greek. In Proceedings of the ESSLLI 03 -15th European Summer School in Logic Language and Information, Workshop "The Meaning and Implementation of Discourse Particles", Stede, M. and Zeevat, H. (eds), Chapter 6, (pp.41--48), Vienna, Austria.

Bailly, G., Banga, E., Monaghan, A., Rank, E. (2000). The Cost258 Signal Generation Test Array. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000) (pp.651--654), Athens, Greece.

CIMWOS project (2000-2003). Details available at: www.xanthi.ilsp.gr/cimwos/

Dutoit, T., Bagein, M., Malfrère, F., Pagel, V., Ruelle, A., Tounsi, N. and Wynsberghe, D. (2000). EULER, An Open Generic, Multi-lingual and Multi-platform Text-to-Speech System. In Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000) (pp.563--566), Athens, Greece.

ERMIS project (2002-2004). Details available at: http://www.image.ntua.gr/ermis

Hoffmann, L. (1989). Textoptimierung im Beispiel "Grammatik": Ein Blick aus der Werkstatt. In G. Antos, G. Augst (Eds.), Textoptimierung. Das Verstaenndlicher-machen von Texten als linguistisches, psychologisches und praktisches Problem (pp.52- -69). Frankfurt am Main.

Kelz, H. P. (ed) (1983). Fachsprache 1. Sprachanalyse und Vermittlungsmethoden. Bonn.

Monrad-Krohn, GH: *Dysprosdy or altered 'melody of language'*. Brain 70, (pp. 405--415), 1948.

Scherer, K. (2000). Emotion Effects on Voice and Speech: Paradigms and Approaches to Evaluation. In Proceedings of the ISCA Workshop on Speech and Emotion, Belfast.

Wardhaugh, R. (1992). Introduction to Sociolinguistics. Oxford: Blackwell.