Workshop on

Compiling and Processing Spoken Language Corpora

Preliminary programme

14.30 Opening

Session 1: Corpus compilation and (orthographic) transcription

14.35 The Corpus of Spoken Israeli Hebrew (CoSIH); Phase I: The Pilot Study
- Shlomo Izre'el and Giora Rahav, Tel-Aviv University

15.00 The ICSI Meeting Corpus: Near-field and far-field, multi-channel transcriptions for speech and language researchers
- Jane Edwards, International Computer Science Institute, and Institute of Cognitive Studies, UC Berkeley

15.25 Orality and difficulties in the transcription of a spoken corpus
- Ana González Ledesma, Guillermo De la Madrid Heitzmann, Manuel Alcántara Plá, Raúl De la Torre Cuesta, and Antonio Moreno-Sandoval, Universidad Autónoma de Madrid

15.50 Processing spoken language data: The BASE experience
- Sarah Creer and Paul Thompson, University of Reading

16.15 – 16.45 Coffee break

Session 2: Corpus annotation

16.45 A "toolbox" for tagging the Spanish C-ORAL-ROM corpus
- José Guirao and Antonio Moreno-Sandoval, University of Granada and Universidad Autónoma de Madrid

17.10 Towards the creation of an electronic corpus to study directionality in simultaneous interpreting
- Claudio Bendazzoli, Cristina Monti, Annalisa Sandrelli, Mariachiara Russo, Marco Baroni, Silvia Bernardini, Gabriela Mack, Elio Ballardini and Peter Mead, University of Bologna

17.35 Developing a dialogue act coding scheme: An experience of annotating the Estonian Dialogue Corpus
- Tiit Hennoste, Mare Koit, Andriela Rääbis, and Maret Valdisoo, University of Tartu

18.00 – 18.15 Short (15-minute) break

Session 3: Extending corpus parameters

18.15 WinPitch Corpus. A text to speech analysis and alignment tool for large multimodal corpora
- Philippe Martin, Université Paris 7

18.40 Automatic annotation of speech corpora for prosodic prominence
- Fabio Tamburini and Carlo Caini, University of Bologna

19.05 Towards dynamic corpora
- Daan Broeder, Hennie Brugman, Nelleke Oostdijk, and Peter Wittenburg, Max Planck Institute for Psycholinguistics and University of Nijmegen

19.30 Closing remarks

Abstracts

- The Corpus of Spoken Israeli Hebrew (*CoSIH*); Phase I: The Pilot Study

  The Corpus of Spoken Israeli Hebrew (CoSIH) is, to the best of our knowledge, the first corpus designed to integrate both demographic and contextual variables in its compilation of texts. The suggested design is culturally dependent to suit the structure of the Israeli Hebrew speech community, yet the principles governing this design are such that they would service study of many other speech communities, to the extent that the design itself may be employed in the compilation of other language corpora with the necessary, culture-dependent modifications. A detailed description of the design can be found in Izre'el, Hary & Rahav (2001).

  In the paper offered for the workshop, we describe the pilot study of *CoSIH*, its procedures and some of its lessons. The results of the pilot study will bring about some changes in the final model of *CoSIH* and in some procedural strategies. We will address a few of the key issues involved in the construction of the corpus in order to achieve the analytical model we have designed. These are: (1) Demographic sampling and recruiting informants; (2) Evaluation of sequential longitudinal recording: technical matters and ethical issues; (3) Contextual sampling: long- and short-term time sampling, speech sampling; (4) The concept of 'cell'. Lastly, the issue of transcription and annotation will be addressed briefly.

- The ICSI Meeting Corpus: Near-field and far-field, multi-channel transcriptions for speech and language researchers

  The recently-completed ICSI Meeting Corpus is available through the LDC. It comprises audio and transcripts of 75 research meetings, ranging in size from 3 to 10 people, with an average of 6 people. The meetings were recorded by means of both close-talking (headset or lapel) microphones and far-field (table-top) microphones. The close-talking microphones (headsets and lapel) enable separation of each person's audible activities from those of every other participant. The far-field microphones provide a view of the meeting as a whole. The transcripts preserve words and other communicative phenomena, displayed in musical score format, time-synchronized to the digitized audio recordings. The corpus is intended as a resource for both speech researchers and language researchers. This paper

describes the methods used to prepare the corpus, some interesting challenges and solutions, and the benefits of using both close-talking and far-field microphones.

- Orality and difficulties in the transcription of a spoken corpus

  This paper analyses the effects of certain oral features on the process of transcription of spontaneous speech recordings. On the basis of the statistical analysis of the data obtained from the C-ORAL-ROM corpus, it will be shown empirically that transcription difficulties vary according to the communicative situation, the degree of formality and the number of participants.

- Processing spoken language data: The BASE experience

  Transcription and mark-up of spoken language data should ideally present as accurate, full and impartial a representation of the original speech event as possible, but processing of the data record is subject to a number of compromises between the pull of competing forces, such as the demand for user readability along with computer readability, and the requirement (for purposes of interchangeability) for conformity to existing standards vs the accurate description of the particularities of the data. This paper presents problems that we have encountered during the process of creating a corpus of orthographically transcribed spoken language data for the British Academic Spoken English corpus. Limitations in the recommendations of the TEI Guidelines are also reviewed.

- A “toolbox” for tagging the Spanish C-ORAL-ROM corpus

  The goals of this paper are to present the tagging procedure for a Spanish spoken corpus, and to show a tool developed for helping human annotators in the process. Some tagging problems especially relevant in spoken corpora, although found also in written texts, will be introduced first.
  The paper will summarize the experience of the group in tagging one of the currently largest spontaneous speech corpora (over 300.000 transcribed words).

- Towards the creation of an electronic corpus to study directionality in simultaneous interpreting

  Spoken corpora have long been awaited in the field of simultaneous interpreting studies. Small scale attempts have provided so far a variety of theories and results which need scientific validation. Our research project aims at contributing to the creation of an electronic parallel corpus which comprises source and target texts in different languages. The following paper presents the initial steps undertaken in this respect and the analysis to be carried out at different levels.

- Developing a dialogue act coding scheme: An experience of annotating the Estonian Dialogue Corpus

  The paper gives an overview of the dialogue act coding scheme that we are developing with the goal to annotate the Estonian dialogue corpus. Our primary task is to analyze Estonian spoken dialogues with the further aim to model human-computer interaction in Estonian. We have studied various coding schemes and tried

to take over the best properties of these schemes. The paper describes our experience of analyzing and annotating the Estonian dialogue corpus.

- WinPitch Corpus. A text to speech analysis and alignment tool for large multimodal corpora

  WinPitch Corpus is an innovative software program for computer-aided alignment of large multimodal corpora. It provides a method for easy and precise selection of alignment units, ranging from syllable to whole sentences in a hierarchical storing system of aligned data. The method is based on the ability to link visually a target with the perception of corresponding speech sound played back. Listening to slower speech, an operator is able to select with a mouse click a segment of text corresponding to the speech sound perceived, and generate by this action bidirectional speech-text pointers defining the alignment. This method has the advantage on emerging automatic processes to be effective even for poor quality speech recordings, or in case of speakers' voice overlap. The software handles multimedia files and is capable to display the corresponding video streams at slower speed.

- Automatic annotation of speech corpora for prosodic prominence

  This paper presents a study on the automatic detection of prosodic prominence in continuous speech, with particular reference to American English, but with good prospects of application to other languages. Perceptual prosodic prominence is supported by two different prosodic features: pitch accent and stress. Pitch accent is acoustically connected with fundamental frequency (F0) movements and overall syllable energy, whereas stress exhibits a strong correlation with syllable nuclei duration and mid-to-high-frequency emphasis. This paper shows that a careful measurement of these acoustic parameters, as well as the identification of their connection to prosodic phenomena, makes it possible to build automatic systems capable of identifying prominent syllables in utterances with performance comparable with the inter-human agreement reported in the literature without using any kind of information apart the acoustic parameters derived directly from speech waveforms.

- Towards dynamic corpora

  In this paper, the idea of a 'Dynamic Corpus Environment' is taken up. After we specify and discuss the functional design of such an environment, the idea is further elaborated upon by illustrating its implementation as it is envisaged, for example, for the Spoken Dutch Corpus and the DOBES Corpus.