Workshop on

COMPILING AND PROCESSING SPOKEN LANGUAGE CORPORA

http://lands.let.kun.nl/CPSLC/

Centro Cultural de Belem, Lisbon, Portugal
24 th May 2004

Workshop to be held in conjunction with
the 4th International Conference on Language Resources and Evaluation (LREC 2004)
Main conference: 26-27-28 May 2004
http://www.lrec-conf.org/lrec2004/

Aim

The aim of the workshop is to bring together people working on the development (compilation and processing) of spoken language corpora.* The workshop will provide participants with the opportunity to exchange views and share experiences. Moreover, the workshop is instrumental in taking stock of and evaluating the present state-of-the-art. The workshop thus aims to contribute to the development of a future roadmap that will guide the development of standards, tools, etc. for use with spoken language corpora.

*The term 'spoken language corpora' is used here to distinguish such corpora from speech corpora or speech databases: speech corpora are collections of spoken data that are typically recorded for specific purposes by specific users (speech corpora/databases such as SpeechDat Car that are used for developing consumer applications). Usually such databases lack the richness of linguistic annations that is pursued for spoken language corpora.

Background and motivation

Despite the wide experience gained in the compilation of written language corpora, working with spoken language data is not immediately straightforward as spoken language involves many novel aspects that need to be taken care of. The fact that spoken language is transient is sometimes offered as an explanation for why it is more difficult to collect spoken data than it is to compile a corpus of written data. However, it is not just the capturing of data that is anything but trivial. Once the (audio) data have been collected and stored, the next step is to produce some kind of transcript (whether orthographic or phonetic). Further annotations such as POS tagging, lemmatisation, syntactic annotation, and prosodic annotation may then build upon this transcription. Among the problems encountered in the processing of spoken language data are the following:

- There is as yet little experience with the large scale transcription of spoken language data. Procedures and guidelines must be developed, and tools implemented.

- Well-established practices that have originated from working on written language corpora do not hold up when trying to cope with the idiosyncracies of the spoken language. This is true for all levels of linguistic annotation. Annotation schemes need to be reconsidered and tools must be adapted.

- In so far as standards have emerged (eg CES), they need to be adapted in order to be able to cater for the needs of spoken language corpora.

- By their very nature, spoken language corpora bring together speech and language technologists and linguists from various backgrounds. Ideally, such corpora should address the needs of all these different user groups. Often, however, there is a conflict of interest. For example, the quality of recordings of spontaneous conversations in noisy environments although highly interesting and worthwhile from a linguistic perspective will prove too poor to be of any use to someone doing research into speech recognition.

Workshop topics

Topics of interest include orthographic transcription, phonetic transcription, prosodic annotation, segmentation, POS tagging and lemmatisation, parsing, and discourse analysis. Contributions on the development and implementation of standards or guidelines for spoken language corpora (annotation schemes, meta-data descriptions) are also invited, as are contributions describing software for the exploitation of spoken language corpora.

Format of the Workshop

The workshop will comprise of oral presentations of previously submitted papers that went through a double peer review process. The proceedings of the workshop will be published by the local organising committee.

Important dates

24th January 2004	Deadline for submission of (full) papers
1st March 2004	Notification of acceptance and preliminary programme
21st March 2004	Deadline for submission of final versions of accepted papers for the proceedings
3rd April 2004	Definitive programme
24th May 2004	Workshop

Submissions

Prospective authors are invited to submit papers for oral presentation. Only full papers in English will be accepted, and the length of the paper should not exceed 6000 words (or the equivalent in space for diagrams). Submissions in MS Word, Postscript, PDF or RTF should be submitted through the workshop website: http://lands.let.kun.nl/CPSLC/

Registration

Workshop participants need to register through the LREC website: http://www.lrec-conf.org/lrec2004/
The fee for this half-day workshop is 50 Euro for conference participants and 85 for others and includes a coffee break and the workshop proceedings.

Organising committee

Nelleke OOSTDIJK, University of Nijmegen
Gjert KRISTOFFERSEN, University of Bergen
Geoffrey SAMPSON, University of Sussex

Programme committee (provisional)

Daan BROEDER	Max Planck Institute
Emanuela CRESTI	University of Florence
Gjert KRISTOFFERSEN	University of Bergen
Tony MCENERY	University of Lancaster
Nelleke OOSTDIJK	University of Nijmegen
Pavel IRCING	University of Western Bohemia
Geoffrey SAMPSON	University of Sussex
Antonio Moreno SANDOVAL	University of Madrid
Jean VERÓNIS	Université de Provence