MAMCO: a Maltese Multimodal Corpus

Patrizia Paggio, Luke Galea and Alexandra Vella
University of Malta
Institute of Linguistics
patrizia.paggio@um.edu.mt, lgal.2812@gmail.com, alexandra.vella@um.edu.mt

Multimodal corpora are becoming increasingly important for the study of communication in different languages. MAMCO is developing the first multimodal resource involving Maltese conversational data. This corpus will pave the way for novel empirical work on gesture and dialogue in Maltese. It will also be used for computational work on multimodal analysis and generation.

The corpus consists of twelve video-recorded first encounter conversations between pairs of Maltese speakers. Twelve speakers participated (6 females and 6 males), each taking part in two different short conversations. The setting and general organisation replicate those used in the Nordic NOMCO corpus (Paggio et al., 2010), where similar corpora have been collected for Danish, Swedish, Estonian, Finnish and Chinese. Participants were instructed to engage in the type of conversation one would have on first meeting someone. They were free to choose their own topics of conversation. In order to assess the naturalness of the conversations, a post-experiment questionnaire was used in which subjects were required to assess each interaction with scores from 1 to 5 along various parameters having to do with how comfortable they had felt during the conversations. For most parameters, the scores fall between 3.5 and 4.5, indicating that the interactions were judged by the participants themselves as reasonably natural.

The annotation of the corpus is underway. The spoken data have been orthographically transcribed using Praat (Boersma and Weenink, 2009) and following the guidelines from Vella et al. (2010). Work on the annotation of head movements is also planned following the MUMIN coding scheme (Alwood et al 2007). Preliminary studies of different aspects of the corpus have been conducted. They include automatic phonetic segmentation, an analysis of overlaps, automatic annotation of head movements and a study of lengthening as a discourse strategy.

References

Allwood, J., L. Cerrato, K. Jokinen, C. Navarretta, and P. Paggio (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. In J.-C. Martin et al. (dds.), Multimodal Corpora for Modelling Human Multimodal Behaviour, Volume 41 of Special issue of the International Journal of Language Resources and Evaluation, pp. 273–287. Springer.

Boersma, P. and D. Weenink (2009). Praat: doing phonetics by computer (Version 5.1.05)

Paggio, P., J. Allwood, E. Ahlsén, K. Jokinen and C. Navarretta (2010). The NOMCO Multimodal Nordic Resource - Goals and Characteristics, in Calzolari et al. (eds.) *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10),* pp. 2968–2974, Valletta, Malta.

Vella, A., Chetcuti, F., Grech, S. & M. Spagnol (2010). Integrating annotated spoken Maltese data into corpora of written Maltese, in *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC ’10)*, Workshop on Language Resources and Human Language Technologies for Semitic Languages, pp. 83-90, Valletta, Malta.