Lisa Brunetti (Aix-en-Provence, France), Stefan Bott, Joan Costa, Enric Vallduví (Barcelona, Spain)

A multilingual annotated corpus for the study of Information Structure

Abstract

An annotated speech corpus in Catalan, Italian, Spanish, English, and German is presented. The aim of the corpus compilation is to create an empirical resource for a comparative study of Information Structure (IS).

Description. A total of 68 speakers were asked to tell a story by looking at the pictures of three text-less picture story books by Mercer Meyer (cf. Berman/Slobin 1994, Strömqvist/ Verhoven 2004, and references quoted there). The participants were mostly university students. The result is 222 narrations of about 2-9 minutes each (a total of about 16 hours of speech). The recordings are transcribed with an orthographic transcription. Transcriptions and annotations of some selected high quality recordings have been aligned to the acoustic signal stream using the program PRAAT (Boersma/Weenink 2009) and its specific format (cf. Bertrand et al. 2008).

Annotation. We propose an original annotation of non-canonical constructions (NCCs) for the Romance subgroup, namely of syntactically/prosodically marked structures that are used to express informational categories such as topic, focus, contrast. The list of NCCs to be annotated is chosen on the basis of our knowledge of the typical NCCs of these languages: left/right dislocations (with or without resumption of a clitic pronoun), cleft and pseudocleft clauses, subject inversion (with or without deaccenting of the subject), fronting of focused elements, etc. Given that these are null-subject languages, the absence of lexical/full-pronoun subjects is also annotated.

The analysis of NCCs in context is extremely useful for the study of IS, as they show *explicitly* what IS strategy the speaker uses within a specific discourse context. Despite their importance, only one example of this kind of annotation is available in the literature, to the best of our knowledge: the MULI corpus of written German (Baumann 2006).Therefore, our corpus provides a so-far missing empirical resource, which will enhance the research on IS based on quantitative analysis of sentences in real context.

Exploitation. The annotation is a useful tool for a comparative description of IS strategies in (Romance) languages with very similar linguistic potential, in particular with respect to the interaction 'IS-syntax' and 'IS-prosody' (via the alignment to the acoustic signal). The annotation is also useful for the study of the interaction 'IS-discourse' (see Mayol 2008).

A survey of the difference in frequency and use of NCCs in the three languages is given. For instance, we show that these languages differ in their strategies to hide the agent of the event. Passives are largely used in Italian, while arbitrary subjects are more common in Spanish. This is presumably connected to a larger use of left dislocations that we witness in Spanish than in Italian and Catalan. These latter languages, on the other hand, exploit **right** dislocation constructions more largely than Spanish. These and similar data allow us to make generalizations concerning the degree of transparency of these languages in their linguistic representation of IS (cf. Leonetti 2008).

References

Baumann, S. (2006), 'Information Structure and Prosody: Linguistic Categories for Spoken Language Annotation'. In: S. Sudhoff et al. (eds.), *Methods in Empirical Prosody Research (Language, Context and Cognition 3)*, de Gruyter.

Berman, R. A., D. I. Slobin (eds.) (1994), *Relating events in narrative: A crosslinguistic developmental study*, Lawrence Erlbaum Associates.

Bertrand, R., P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, S. Rauzy (2008), 'Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle'. In: *Traitement Automatique des Langues*, 49, 3.

Boersma, P., Weenink, D. (2009), 'Praat: doing phonetics by computer' (Version 5.0.47) [Computer program]. Retrieved January 21, 2009, from http://www.praat.org/

Leonetti, M. (2008), 'Alcune differenze tra spagnolo e italiano relative alla struttura informativa'. Paper presented at *Convegno dell'Associazione Internazionale dei Professori d'Italiano*, Oviedo, Sept. 3-6, 2008.

Mayol, L. (2008), 'Overt pronouns in Catalan: information, discourse and strategy', Dissertation proposal. Advised by Robin Clark. University of Pennsylvania.

Strömqvist, S., L. T. Verhoven (eds.) (2004), *Relating events in narrative, Vol. 2: Typological and contextual perspectives*. Lawrence Erlbaum Associate.