Typological Data on Information Structure

Caroline Féry, Stavros Skopeteas & Ruben Stoel

University of Potsdam

fery@uni-potsdam.de

Design of the questionnaire

In the first part of the talk, we will present the questionnaire (QUIS) we have elaborated for the query of linguistic correlates of information structure, in the framework of project D2[1] of the SFB 632 in Potsdam. The questionnaire leans on a rich tradition of systematic queries on typological and dialectal differences (see for instance Dahl 1985; Bybee & Dahl 1989; Dahl 2000 for temporal, aspectual and modal categories) and it aims at systematic rigor (see the Eurotyp project as documented in Veselovská 2000; Vos & Veselovská 2001). The collected data are organized in a database. The query, designed by project D1 of the same SFB, allows users to access the data from different gates: language, grammar, information structure and tasks being the most important ones. Carefully transcribed and annotated sound files (ESMERALDA), description of the tasks used to elicit the data, as well as the way they have been collected, are also part of the database. Once ready, the database will be made available to the linguistic community: access to our data will be unlimited. Other linguistic databases are for example the „Penn TreeBank“ (Marcus, Santorini & Marcinkiewicz 1993) and the Saarbrücken „NEGRA“ corpus (Skut et al. 1997).

QUIS comprises several parts: a set of questions about grammar, sentences to be translated and summaries to be prepared about the grammatical correlates of information structure in the language under consideration. The central and most original part of QUIS consists of 30 tasks with different information structural contents, aimed at eliciting spontaneous speech. Wide and narrow focus, contrastive focus, double foci, partial topic, implicational topic, contrastive topic, whole event, new/given partitioning, bridging are examples of the phenomena elicited. Some tasks also vary the animacy of the participants (agents and patients), and/or their visibility. Some others bear on the expression of spatial relationships, and on quantification. The tasks are mostly based on non-verbal or partly verbal material: pictures, map tasks, games, movies, etc are widely used. Some further tasks are conventional question-answer pairs, depending on the kind of data to be elicited. The

[1] In collaboration with Gisbert Fanselow, Ines Fiedler, Manfred Krifka and Anne Schwarz.

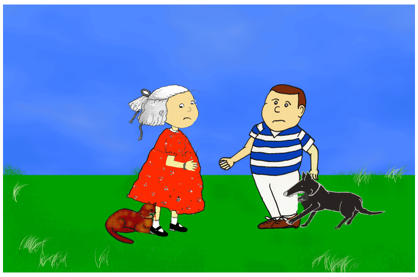
questionnaire can be used in typologically different languages, in different cultures, with minimal preliminary adaptations. Crucially, all tasks are neutral as to the linguistic devices to be elicited. Not a single one is specifically phonological, morphological or syntactic, though there are of course tasks which tend to elicit devices readily expressed by syntactic means (like passive vs. active or pronominalization) or by phonological means (like contrastive focus on an adjective).

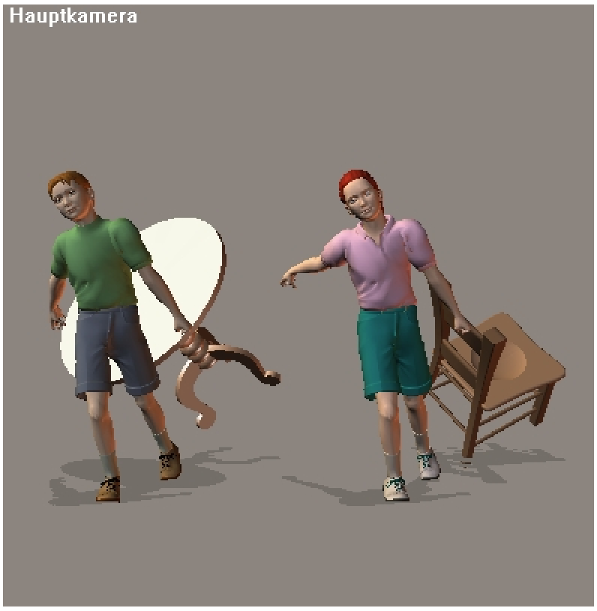
In the first phase of the project, extensive data from 15 partly genetically different languages are gathered: German, Dutch, English, French, Greek, Hungarian, Georgian, Japanese, Mandarin, Prinmi, Konkani, Mawng, Niue, Terribe and Yucatec Maya. The well-studied languages among them allow us to verify the usefulness of the tasks. In particular, German has been used as a test language to refine many of the tasks.

Task illustration

Methodologically, all tasks allow an in-depth investigation of the information structural pattern they elicit. The 30 non-verbal or partly verbal tasks are conceived like as many experimental set-ups which can be applied to a large population of subjects, so that results coming from each language and each task can deliver statistically relevant results.

The task used as an illustration is bearing on elicitation of double foci. The aim of this task is to inspect the grammatical devices used to express double foci cross-linguistically, as compared to the devices used in situation where only one focus is needed. The data set is induced through questions on visual material, as those shown in Fig.1. Most of our pictures have been made using the software POSER, but some (like the first one in Fig.1), have been drawn by a professional designer.





The experimental conditions in the double foci task are summed up in table 1.

Condition	verbal stimulus	visual stimulus
A	Double wh-question (list)	Who is biting who?
B	Agent focus, double question	Who is biting the boy and who is biting the girl?
C	Patient focus, double question	Whom is the dog biting and whom is the cat biting?
D	All new (double action)	What happens?
E	Double wh-question (single pair)	Who is biting whom?
F	Agent focus, single question	Who is biting the boy?
G	Patient focus, single question	Whom is the dog biting?
H	All new (single action)	What happens?

Like all other experiments, this experiment has been designed factorially: Each experimental condition is implemented in 8 items and each informant is confronted with each item once and each condition once (in randomized order), hence producing in sum 8 sentences (one per condition). Up to this point, the experiment has been performed in its entirety in 4 languages: English, Georgian, German, and Greek, the results of which we will sum up in our talk. 16 native speakers in each language have participated in the experiment, which results in a total of 128 sentences per language.

3 Experimental results on answers to double wh-questions

The recorded data are examined from different points of view. First in a language-specific way and second in a cross-linguistic perspective. The main devices used in each condition are then extracted and summed up. Only representative examples are selected to be annotated in their entirety and only those enter the ‘core’ of the database.

Prosody and syntax are especially important in the data considered here. Number, excursion and direction of accents, phrasing and tonal scaling are the most relevant prosodic criteria. Passivization, word order, ellipsis and the use of morphological markers are the morpho-syntactic factors which vary most in these data.

Georgian is making use of word order change more than German, Greek and English. In conditions in which the informants see two pictures and answer only to one question, grammatical devices are used which are reminiscent of implicational topic.

Statsitical analysis both in a inner-linguistic and in a cross-linguistic perspective will be presented. The importance of studying comparable data on a large scale in order to understand better hwo information structure is realized will be emphasized.

References

Bybee, J. and Ö. Dahl (1989). The Creation of Tense and Aspect Systems in the Languages of the World. Studies in Language, **13**: 51-103.

Dahl, Ö. (1985). Tense and Aspect Systems. Basil Blackwell, Oxford.

Dahl, Ö. (2000). Tense and Aspect in the Languages of Europe. Mouton de Gruyter, Berlin.

Marcus, M., B. Santorini and M. Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics, **19(2)**: 313-330.

Skut, W., B. Krenn, Th. Brants and H. Uszkoreit (1997). An Annotation Scheme for Free Word Order Languages. In: Proc. of the Fifth Conference on Applied Natural Language Processing (ANLP-97), Washington/DC.

Veselovska, L. (2000). Elements of a System. Habilitationsschrift, Olomouc.

Vos, R. and L. Veselovská (2001). Clitics in the Languages of Europe. In: H. Riemsdijk, ed., Clitics in the Languages of Europe.