The SoSaBiEC Corpus:

Social Structure and Bilinguality in Everyday Conversation

Veronika Ries[1], Andy Lücking[2]

[1]Universität Bielefeld, BMBF Projekt Linguistic Networks

[2]Goethe-Universität Frankfurt am Main

E-mail: Veronika.Ries@uni-bielefeld.de, Luecking@em.uni-frankfurt.de

Abstract

The SoSaBiEC corpus is comprised audio recordings of everyday interactions between familiar subjects. Thus, the material the corpus is based on is not gained in task-oriented dialogue under strict experimental control; rather, it is made up of spontaneous conversations. We describe the raw data and the annotations that constitute the corpus. Speech is transcribed at the level of words. Dialogue act oriented codings constitute a functional, qualitative annotation level. The corpus so far provides an empirical basis for studying social aspects of unrestricted language use in a familiar context.

Keywords: bilinguality, social relationships, spontaneous dialogue, annotation

1. Introduction

From the point of view of the methodology of psycholinguistic research on speech production unconstrained responding behavior of participants is problematic: it is known as “the problem of exuberant responding” and it is to be avoided by means of some sort of controlled elicitation in an experimental setting (Bock 1996, p. 407; see also Pickering and Garrod 2004, p. 169). In addition, elicitations are usually bound up with a certain task the participants of the experimental study have to accomplish. Of course, each experimental set-up that obeys to the general “avoid-exuberant-responding”-design and is therefore appropriate to study and test the conditions underlying speech production in a controlled way. However, when studying human-to-human face-to-face dialogue (or multi-logue, in case of more than two interlocutors), elicited communication behavior hinders the unfolding of spontaneous utterances and task-independent dialogue management. Task-oriented dialogue is known to be plan-based (Litman and Allen 1987). The domain knowledge the interlocutors have of the task-domain together with the difference between their current state and the target state (defined in terms of the task to be accomplished) provides a structuring of dialogue states: the way from the current dialogue state to the target state is operationalized as a sequence of sub-tasks, each of these sub-tasks is part of a plan that has to be worked off sequentially in order to reach the target state. Plan-based accounts to dialogue provide a functional account to dialogue and have been successfully applied in computational dialogue systems for, e.g., timetable enquiries (Young and Proctor 1989). At least partly due to the neat status of task-oriented conversational settings, respective study designs have been paradigmatic in linguistic research on dialogue. Task-oriented dialogues, inter alia, pre-determine the following conversational ingredients:

- they define a dialogue goal and thereby a terminal dialogue state;
- they constrain the topics the interlocutors talk about to a high degree (up to move type predictability, modulo repairs etc.);
- they are cooperative rather than competitive;
- the dialogue goal determines the social relationship of the interlocutors (for instance, whether they have equal communicative rights or whether task-knowledge is asymmetrically distributed) and it does so regardless of the actual relationships that might obtain between the interlocutors;

- they are unilingual.

Each of the ingredients above is lacking in spontaneous, everyday conversation. Does this mean that spontaneous, everyday conversations also lack any structure of dialogue management? Answers to this question are in general given on the grounds of armchair theorizing or case studies. The feasibility of empirical approaches is simply hindered by the lack of respective data. The afore-given list can be extended by a further feature, namely the fact that it is easier to gather task-oriented dialogue data in experimental settings than to collect rampant spontaneous dialogue data. We have some spontaneous dialogue data that lack each of the task-based features listed above – see section 2 for a description. We focus on the latter two aspects here, namely social structure and bilingualism. The social dimension of language use, for instance, social deixis, is a well-known fact in pragmatics (Anderson and Keenan 1985; Levinson 2008). The influence of social structure on the structure of lexica has also been reported (Mehler 2008). Yet, there is no account that scales the macroscopic level of language communities down to the microscopic level of dialogue. The data collected in SoSaBiEC aims at exactly this level of granularity of social structure and language structure: how does the social relationship between interlocutors affect the structure of their dialogue lexicon?

A special characteristic of SoSaBiEC is bilingualism. The subjects recorded speak Russian as well as German, and they make use of both languages in one and same dialogue. What dialogical functions performed by the two languages seems to depend at least partly on who the addresses are, that is, on the social relationship between the interlocutors (Ries to appear). This qualitative observation will be operationalized in terms of quantitative analyses that focus on the relationship-dependent, functional use of languages (cf. the outlook given in section 4).

According to the bi-partition of corpora – primary or raw data are coupled with secondary or annotation data (loosely related to Lemnitzer and Zinsmeister 2006, p. 40) – the following two sections describe the data material (section 2) and its annotation (section 3) in terms of functional dialogue acts. In the last section, we sketch some research question we will address by means of SoSaBiEC in the very near future.

2. Primary Data

The primary data are made up of audio recordings of everyday conversations (Ries to appear). The recorded subjects all know each other, most of them are even related. The observations focus on natural language use, and in particular on bilingual language use. The compiled corpus is authentic because the researcher, who recorded, is herself a member of the observed speech community. The speaker gave their consent for recording at any time and without prior notice. So the recordings were taken spontaneously and at real events, such as birthday parties. For recording a digital recorder without microphone was used, so that it was without attracting too much attention. They include telephone calls and face-to-face conversations. The length of the conversations varies from about three minutes up to three hours. Depending on the topic of the conversation the number of the involved speakers differs: from two up to four speakers. In sum, there are about 300 minutes of data material covering six participants. Altogether ten conversations have been recorded. Four conversations have been analysed in detail and annotated because the participant constellation is obvious and definite: the participants come under the category parent-child or sibling. The six participants come from two families, not known to each other. As working basis for the qualitative analysis the recordings were transcribed. By way of illustration, an excerpt of the transcribed data is given:

```
01 F: NAME
   A: guten abend.
   F: hallo?
   A: hallo guten abend
05 F: nabend (.) hallo
   A: na wie gehts bei euch?
   F: gut
   A: gut?
   F: ja.
10 A: na что вы смотрели что к чему тама?
   F: ja а что там?
```

This is a sequence of a telephone call between father F and his daughter A. The conversation starts in German and initiated by daughter A there is an alternation into Russian (line 10). The qualitative analysis showed that through this language switch speaker introduced the first topic of the telephone call and so managed the conversation opening. Results such as the described one are the main content of the annotation.

3. Annotation

The utterances produced by the participants have been transcribed using the Praat tool (http://www.fon.hum.uva.nl/praat/) on the level of orthographic words. That means, that no phonetic features like accent or defective pronunciations are coded. However, spoken language exhibits regularities of its own kind, regularities we accounted for in speech transcription. Most prominently, words that are separated in written language might get fused into phonetic word in spoken language. A common example in German already part of the standard of the language is “zum”, a melting of the preposition “zu” and the dative article “dem”. Meltings of this pattern are much more frequent in spoken German than acknowledged in standard German. The English language knows hard-wired combinations like “I’m” which usually is not resolved to the full-fledged standard form “I am”. The annotation takes care for these demands in providing respective adaptations of annotations to spoken language. In order to reveal the dialogue-related functions performed by the utterances, we employed a dialogue act-based coding of contributions. Here, we follow the ISOCat (www.isocat.org) initiative for standardization of dialogue act annotations outlined by Bunt et al. (2010).

To be able to talk about dialogue-related functions and natural bilingual language use, language alternations regarding their functions and roles in the current discourse were annotated. The important factor annotated is the function of the involved languages and the observed language alternations: That means to annotate each language switch and its meaning on the level of conversation, for example the conversation opening. The differentiation by speakers is crucial for the examination of a connection between language use and social structure. The functional annotation labels have been derived from qualitative, ethnomethodological analyses by an expert researcher. The annotations made by this vey researcher can be regarded as having the privileged status as “gold standard” since part of the expert’s knowledge is not only the pre- and posthistory of the data recorded, but also familiarity with the subjects involved, a kind of knowledge rather exclusive to our expert. However, since the annotation are a compromise between the qualitative and quantitative methods and methodologies that are brought together in this kind of research, we want to assess whether the ethnomethodological, functional annotation can be reproduced to a sufficient degree by other annotators. For this reason, we applied a reliability assessment in term of inter-rater agreement of two raters’ annotations of a subset (one conversation) of the data. We use the agreement coefficient AC1 developed by Gwet (2001). The annotation of dialogue acts result in an AC1 of 0.61, the rating of function result in an AC1 of 0.78. Two observations can be made: firstly, the functional dialogue annotation is reproducible -- an outcome of 0.78 is regarded as "substantial" by Rietveld and van Hout (1993); secondly, the standardised dialogue act annotation scheme tailored for task-oriented dialogues can be applied with less agreement than the functional scheme custom-build to more unconstrained everyday conversations. We take this as further evidence for the validity of the distinction of different types argued for in the introduction.

4. Outlook

So far, we finished data collection and annotation of the subset of SoSaBiEC data that interests us first, namely the data that involve parent-child and sibling dialogues. The next step is to test our undirected hypothesis by means of mapping the annotation data on a variant of the dialogue lexicon model of Mehler, Lücking, and Weiß (2010). This model provides a graph-theoretical framework for classifying dialogue networks according to their structural similarity. Applying such quantitative measure onto mostly qualitative data allows not only to study whether social structure imprints on language structure in human dialogue, but in particular to *measure* if there is a traceable influence at all.

5. Acknowledgments

Funding of this work by the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung) is gratefully acknowledged. We also want to thank Barbara Job and Alexander Mehler for discussion and support.

6. References

Anderson, Stephan R. and Edward L. Keenan (1985). “Deixis”. In: Language Typology and Syntactic Description. Ed. by Timothy Shopen. Vol. III. Cambridge: Cambridge University Press. Chap. 5, pp. 259–308.

Bock, Kathryn (1996). “Language Production: Methods and Methodologies”. In: Psychonomic Bulletin & Review 3.4, pp. 395–421.

Bunt, Harry et al. (May 21, 2010). “Towards an ISO Standard for Dialogue Act Annotation”. In: Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC’10). Ed. by Nicoletta Calzolari (Conference Chair) et al. Valletta, Malta: European Language Resources Association (ELRA).

Cohen, Jacob (1960). “A Coefficient of Agreement for Nominal Scales”. In: Educational and Psychological Measurement 20, pp. 37–46.

Gwet, Kilem (2001). Handbook of Inter-Rater Reliability. Gaithersburg, MD: STATAXIS Publishing Company.

Lemnitzer, Lothar and Heike Zinsmeister (2006). Korpuslinguistik. Eine Einführung. Tübingen: Gunter Narr Verlag.

Levinson, Stephen C. (2008). “Deixis”. In: The Handbook of Pragmatics. Blackwell Publishing Ltd, pp. 97–121.

Litman, Diane J. and James F. Allen (1987). “A plan recognition model for subdialogues in conversations”. In: Cognitive Science 11.2, pp. 163–200.

Mehler, Alexander (Mar. 2008). “On the Impact of Community Structure on SelfOrganizing Lexical Networks”. In: Proceedings of the 7th Evolution of Language Conference (Evolang 2008). Ed. By Andrew D. M. Smith, Kenny Smith, and Ramon Ferrer i Cancho. Barcelona: World Scientific, pp. 227–234.

Mehler, Alexander, Andy Lücking, and Petra Weiß (2010). “A Network Model of Interpersonal Alignment in Dialogue”. In: Entropy 12.6, pp. 1440–1483. doi: 10.3390/e12061440.

Pickering, Martin J. and Simon Garrod (2004). “Toward a Mechanistic Psychology of Dialogue”. In: Behavioral and Brain Sciences 27.2, pp. 169–190.

Ries, Veronika (2011). “da=kommt das=so quer rein. Sprachgebrauch und Spracheinstellungen Russlanddeutscher in Deutschland”. PhD thesis. Universität Bielefeld.

Rietveld, Toni and Roeland van Hout (1993). Statistical Techniques for the Study of Language and Language Behavior. Berlin ; New York: Mouton de Gruyter.

Young, S. J. and C. E. Proctor (1989). “The design and implementation of dialogue control in voice operated database inquiry systems”. In: Computer Speech and Language 3.4, pp. 329–353. doi: 10.1016/0885-2308(89)90002-8.