M. Kuehnast

DEVELOPING A CORPUS FOR CONTRASTIVE STUDIES OF INTERSENTENTIAL ANAPHORA IN CHILD LANGUAGE

1. Introduction

The paper deals with the theoretical and practical issues of creating a new parallel corpus of child speech. The project is based at the Centre for General Linguistics and Typology (ZAS) in Berlin. The research of the team guided by Prof. Dagmar Bittner concentrates on cross-language studies of anaphora resolution in the speech of children acquiring closely related Slavic languages – Bulgarian and Russian – or a typologically different language – German. The research on Russian is conducted by Dr. Natalia Gagarina, on Bulgarian by the author. Dr. Insa Guelzow is in charge of the German data base.

On the basis of cross-language comparison we investigate the acquisition of strategies used for referent disambiguation of intersentential anaphora, aiming at differentiating between processing factors such as memory and attention limitations and the internal organisation of language structures. In order to trace down the effects of semantic, syntactic and pragmatic factors and their interactions in the complex process of anaphora resolution we need specific data reflecting the emergence of intersentential nominal and pronominal reference in production. As groundwork we develop a parallel corpus of Bulgarian, German and Russian narratives featuring a fine grained annotation sensitive to the idiosyncrasies of early child speech.

2. Theoretical issues

Comparative data on Bulgarian, German and Russian highlights elements specific to those languages against the background of general characteristic traits of anaphoric reference. Linguistically motivated properties of available referents such as agenthood, subject or topic status may obtain different cue validity in the investigated

languages. The choice of the data base languages is governed by the expected impact of their typological characteristics on the development of nominal and pronominal intersentential reference in the speech of young children:

Table 1. Selected typological properties of the corpus languages

	German	**Bulgarian**	**Russian**
Word Order	relatively free, SOV	relatively free, SVO	relatively free, SVO
Information structure	subject/agent orientated	pronominal marking of object topics	topic orientated
Case system	4 cases	no nominal cases	6 cases
Subject marking	non pro-drop, expletive Subjects	strong pro-drop	weak pro-drop
Nominal Definiteness	pre-posed definite articles	post-posed definite articles	no definite articles

German, Bulgarian and Russian feature pronominal inventories quite different in size and in the purposes they serve even within pronominal classes of the same type. German uses the opposition between personal and demonstrative pronouns to disambiguate reference to the subject or to the object in the preceding sentence. Russian utilises the subclasses of distance and proximity pronouns for the same purpose, whereas Bulgarian does not employ personal or demonstrative pronouns in this function.

On the other hand, Bulgarian makes extensive use of its system of long and short pronominal forms explicating the information structure of utterances. Object topics are marked by means of pronominal clitic doubling. One main function of this strategy is to compensate for the non-existent case markings which trigger subject-object disambiguation in Russian and German.

Nominal definiteness is the third area in which the chosen languages diverge from each other in great extend. The various ways of expressing nominal definiteness - be it overtly by systems of definite articles (German and Bulgarian) or covertly by means of word

order and verbal aspect (Russian) - influence the anaphoric functions of determiners in those languages.

German, Russian and Bulgarian are expected to diverge with respect to the ranking of particular linguistic properties of anaphoric expressions in the hierarchy determining the accessibility status of possible referents.

3. Practical issues

The corpus comprises 3 data bases each containing data of 180 monolingual Bulgarian, German and Russian children. The data bases are structured by age brackets, 6 months apart. Each age bracket (3;0 – 3;6 – 4;0 – 4;6 – 5;0 years) contains the data of 30 children.

The speech data consists of 2 narratives per child elicited in a picture-story-telling design. Each story contains 6 pictures, as line drowning in black and wait. The make up of the stories triggers different forms of referent tracking, suitable to pursue strategies of anaphoric reference within contexts of varying ambiguity.

The first story we used is the «Cat story» by M. Hickmann[1]. The protagonists of the story (mother bird and baby birds, a cat and a dog) are referred to by nouns of different gender and number in each of the languages. Not all characters appear in every picture.

The second story called «Bird story» was created by our team to insure highest level of comparability between the textual properties of Bulgarian, German and Russian data and to provide for maximal ambiguity of pronominal anaphora referring to the protagonists of the story. The plot of the «Bird story» is reminiscent of the well known fable by Jean de La Fontaine «Raven and Fox»[2]. We decided to use a bird of a neutral shape and colour and replaced the piece of cheese by a fish. At the end the bird succeeds in getting back the fish. The «Bird story» promotes ambiguity between the protagonists, both of them being animate and active actors in plot of the story. Both are suitable

[1] *Hickmann, M.* Children's discourse. Person, space and time across languages. Cambridge Studies in Linguistics, 98, Cambridge, 2003.

[2] Jean de La Fontaine: Fabeln. Stuttgart, 1987.

as main topics of the narrative. Ambiguity is endorsed also on the linguistic side because the nominal expressions coding the potential referents are of the same gender - in German they are masculine nouns *Vogel*, *Fuchs* and *Fisch*, in Russian and Bulgarian they are feminine nouns *птичка / птица, лиса / лисица* and *рыба / риба.*

The children are not presented with all 6 pictures at ones, but see only two pictures at the same time, the previous and the new one. This procedure allows for a differentiation between short and long distance anaphoric relations (within the episode of a given picture and between pictures).

The narratives are recorded by means of a digital video camera. The choice of the recording device was guided by the fact, that children often disambiguate vague expressions by pointing at the intended referent on the picture. Young children often substitute nouns with related meanings (*fox – dog – wolf*) or use morphologically inappropriate pronominal forms (gender, number or case errors).

The narratives are transcribed in CHAT format[3] by native speakers with linguistic training. The transcripts are structured in episodes 1 to 6 according to the pictures of the story. Information on external disambiguation (pointing) is added on a comment line.

Subsequently, the transcripts are tagged manually for relevant morphological, syntactic, semantic and pragmatic information. All nominal and pronominal phrases are coded on a separate line. Each referring expression receives several tags identifying its:

- Syntactic status: S (Subject); O (Object) etc.
- Phrase type: DAN (definite noun); IAN (indefinite noun) PP (personal pronoun); Ø (zero pronoun - Subject drop) etc.

The number and type of tags deviate slightly in the single languages. Due to the specific purposes of the corpus, all anaphoric expressions are provided with tags resolving the properties of their antecedents such as animacy (BL/UBL), syntactic status, and distance relations counted in terms of propositions and episodes (between pictures). The following example is from the «Bird story» told by a 5

[3] *MacWhinney, B.*: The CHILDES Project: Tools for Analyzing Talk. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

years-old Bulgarian girl. English translation is given in brackets.

@G:1 (Picture 1)
**EX1: Ja sega mi кazhi, kakvo stava tuka.*
(Well, tell me now what is going on)
**027: Tuka… tuka edna ptitsa idva i vizhda edna riba i ja vzima.*
(Here… here a bird is coming and is spotting a fish and is taking it.)

%cod: edna ptitsa|S-IAN-BL-NV edna riba|O-IAN-UBL-NV я|O-PP-UBL-V0_AO

@G:2 (Picture 2)

**EX1: A tuka? (And here?)*
**027: togava ptitsata ja vze i se kachi na dyrvoto, no togava lisitsata ja vid'a i kaza na ptitsata:*
«Daj mi… daj mi tazi riba!», ama ptitsata ne iskashe.
(Then the bird took it and got on the tree, but the fox saw it and spoke to the bird «Give me, give me this fish!» but the bird didn't want to.)

%cod: ptitsata|S-DAN-BL-V1_1_AS ja|O-PP-UBL-V1_1_AO na dyrvoto|P-DANM-UBL-NV lisitsata |S-DAN-BL-NV ja|O-PP-UBL-V1_AX na ptitsata|P-DANM-BL-V1_AS ми|O-PP-BL-V1_AS tazi riba |O-DP:N-UBL-V4_AO ptitsata|S-DAN-BL-V1_AS

4. Conclusion

The corpus design provides a reliable source for statistic computations as empirical base for linguistic inferences concerning the acquisition of intersentential anaphoric reference in child language, also from a typological perspective. The corpus is to be extended continuously by the data of school children and adult speakers of German, Bulgarian and Russian.