Word-order variation: Why corpus and judgment data do not go hand in hand!

Markus Bader and Jana Häussler

Department of Linguistics, University of Konstanz

(markus.bader|jana.haeussler)@uni-konstanz.de

1 Introduction

We will present results from an ongoing study that combines corpus analyses with on-line experiments in order to investigate the factors determining word-order variation. In German, the language of our studies, most sentences have the word order subject-before-object (SO), but the reverse order (OS) is also possible. Two main linguistic sources for OS have been identified. First, some verbs (e.g. unaccusative verbs and certain psych-verbs) specify OS as the canonical order among subject and object (base-generated OS-order). Second, verbs that normally go with SO-order might occur with OS-order for discourse reasons, involving notions like information structure and topicality; these cases have to be further subdivided into scrambling sentences (OS within the so-called 'middlefield', the part of the sentence between complementizer and clause-final verb) and topicalization sentences (object fronted to clause-initial position, the so-called 'prefield').[1] In addition to, or instead of, these two sources, length considerations have been claimed to play a role, favoring the order OS when the object is shorter than the subject (Hawkins, 2004).

2 Data

Since we were also interested in processes of syntactic ambiguity resolution, all studies investigated sentences with an object introduced by the definite article *den*, which is ambiguous between accusative and dative. For reasons of space, only results for unambiguous sentences will be considered here.

(i) *Experimental findings*. Our comprehension experiments investigated sentences containing a *den*-object in different positions, and with either action or psych verbs. An

[1] We use the terms 'base-generated', 'scrambling', and 'topicalization' in a purely descriptive manner here.

illustrative example of a psych-verb sentence with subject-before-object (SO) word order is shown in (1) ((1) contains a dative object; a corresponding accusative verb is *beeindrucken* 'to impress').

(1) Ich glaube, dass der Vortrag den beiden Studenten gefallen hat.
I believe that the talk the-DAT both students pleased has
'I believe that the talk pleased the two students.'

The two OS-sentences corresponding to (1) are shown in (2). In (2-a), subject and object are again both within the middelfield. In (2-b), the object has been fronted.

(2) a. Ich glaube, dass den beiden Studenten der Vortrag gefallen hat
I believe that the-DAT both students the talk pleased has
'I believe that the talk pleased the two students.'
b. Den beiden Studenten hat der Vortrag gefallen
The-DAT both students has the talk pleased
'The talk pleased the two students.'

The same kind of word-order variation as in (1) and (2) is also possible with sentences containing action verbs (e.g. dative *helfen* 'to help' versus accusative *unterstützen* 'to support').

Comprehension data come from a series of experiments using the method of speeded grammaticality judgments. This method has been shown to be quite sensitive to pure processing effects (sentence complexity, ambiguity resolution), as well as to subtle differences between different types of grammatical sentences (Bader and Bayer, 2006). Sentences were rapidly presented word by word in the center of a computer screen. Immediately after the last word, participants had to judge the grammaticality of the sentence.

For SO-sentences, no differences between accusative and dative were obtained, independent of construction or verb type. For OS-sentences with dative object, the main findings were that scrambling OS-sentences had a slight disadvantage in comparison to SO-sentences, whereas both base-generated OS-sentences and topicalization OS-sentences were indistinguishable from corresponding SO-sentences. OS-sentences with accusative object, in contrast, showed a substantial disadvantage in the condition scrambling, no disadvantage in the condition base-generation, and a slight disadvantage in the condition topicalization.

(ii) *Corpus data*. So far, three sentence sets randomly sampled from the newspaper part of the COSMAS-System (IDS, Mannheim) were analysed: Set1 = 1210 embedded clauses with a *den*-object, unconstrained by position; this sentence set contains 86% SO-sentencess and 14% OS-sentences; Set2 = 824 embedded clauses with the

den-object immediately following the clause-initial complementizer; Set3 = 804 main clauses with a *den*-object in clause-initial position. The sentence sets were annotated for case, animacy, definiteness, pronominality, and length of subject and object, with length defined as number of words. A main finding of the corpus study is that the ratio of accusative to dative sentences is strongly dependent on construction type: For SO-sentences, 82.6% sentences with accusative object versus 17.4% sentences with dative object; for OS middlefield sentences, 6.3% sentences with accusative object versus 93.7% sentences with dative object; for OS prefield sentences, 74.8% sentences with accusative object versus 25.2% sentences with dative object.

With regard to the reasons for using OS-order, our corpus data do not provide evidence for length being an important factor. Instead, our data give rise to the two generalizations in (3) and (4).

(3) OS-word order in the middlefield is almost exclusively used for argument-structure reasons, for both dative and accusative objects (somewhat stronger so for dative objects).

(4) For objects in the prefield, accusative and dative behave differently:
   a. Sentences with fronted accusative objects have the same lexical-semantic properties as corresponding SO-sentences; the OS-order for them seems to be discourse conditioned.
   b. Sentences with fronted dative objects have similar properties as base-generated OS-sentences; the OS-order for them seems to be argument-structure driven.

There are several pieces of evidence for these two generalizations which we will discuss in more detail in our presentation. Here we can give just a single example. We fitted a logistic regression to Set1 which contains both SO- and OS-sentences, with the NP properties animacy, definiteness, pronominality and length as predictor variables. Of these variables, all were significant with the exception of length. The estimated parameters (with predicted p-values of .5 and greater converted to SO and otherwise to OS) correctly predicted the word order of 94% sentences. When the same parameters were applied to the other two sentence sets, correct prediction rates were 84% for OS-middlefield sentences but only 18% for OS-prefield sentences. Given that parameter estimation was based on middlefield sentences, this clearly shows that the use of OS-order in the middlefield depends on different factors than the use of OS-order with topicalized objects.

(iii) *Comparison between corpus and experimental data*. Comparing the on-line judgment data with the corpus data shows that some sentence types which are rare within a corpus can still be easily judged as fully grammatical. Thus, there is no simple coupling of corpus frequency with either grammaticality or comprehension difficulty, as

claimed by, e.g., MacDonald and Christiansen (2002). Instead, the following two conclusions can be derived. First, when an OS-order is lexically licensed by the verb's argument structure, judgments are as high as for corresponding SO-sentences. Second, difficulties are only observed when the OS-order is not lexically but discourse licensed, with the degree of the difficulty modulated by both the case of the object and the particular structural configuration (scrambling versus topicalization). This modulation is partly reflected by the corpus data: Scrambling is the most difficult condition, and scrambling is almost absent from the corpus (observed instances of OS-sentences in the middlefield mostly being instances of base-generation).

3 Discussion

The discrepancies between corpus data and experimental data underline the necessecity of approaching the topic of word-order variation with different empirical methods. While the corpus data help to identify the variables determining word-order variation, the experimental data show that there is no simple relationship between corpus frequency and perceived grammaticality. We will show how a model of the human parsing mechanism built on prior experimental results can account for the judgment results presented here if two assumption are made: (i) the parser computes the focus potential of a sentence based on the argument structure associated with the verb, and (ii) judgments involving word order variation are mainly determined by focus-structure markedness.

References

Bader, M. and J. Bayer (2006). Case and linking in language comprehension - evidence from German. Springer, Heidelberg.

Hawkins, J. A. (2004). Efficiency and complexity in grammars. Oxford University Press, Oxford.

MacDonald, M. C. and M. H. Christiansen (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). Psychological Review, **109**:35–54.