

Predicting subjects' argument structure preferences: A matter of adequate corpus data

Sandra Pappert¹, Johannes Schließer¹, Dirk P. Janssen², & Thomas Pechmann¹

http://www.uni-leipzig.de/~parsing/

Introduction

1 Behavioral data

In German sentences with the main verb in final position, argument-specific information can modulate the availability of argument structures (Friederici & Frisch, 2000).

- In this study, we investigate whether factors that influence word order preferences (e.g. Rösler et al., 1998; Featherston, 2005) have an impact on argument structure preferences:

Dative before Accusative and Definite before Indefinite.

2 Corpus data

Corpus data may serve as a predictor of behavioral data (e.g., Lapata, Keller, & Schulte im Walde, 2001).

- 2-argument structures are more frequent than 3-argument structures,
- Datives tend to precede accusatives (Kempen & Harbusch, 2003, 2004).
- Animacy influences word order, too, but data on 3-argument structures are sparse.
- Definiteness does not correlate with word order (Weber & Müller, 2004).

Behavioral data

3 Sentence completion

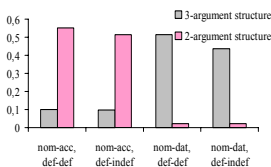
Method

Factors: Case and Definiteness of argument2

nom-acc, def-def vs. nom-acc, def-indef vs. nom-dat, def-def vs. nom-dat, def-indef

Der Doktor wird den
einen
dem
einem Krankenpfleger ... *the_{nom} doctor will a/the_{acc/dat} (male) nurse ...*

Proportion of completions



Hierarchical loglinear model

Z (Case*Argument Structure) = 18
 Z (Argument Structure) = 11
 Z (Case) = 9
 Other Z-values < 1

Discussion

The interaction of Case and Argument Structure (2- or 3-) is not predicted by the available syntactic frequency data. Definiteness has no significant effect.

4 Self-paced reading

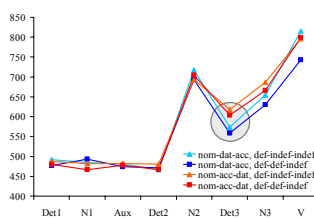
Method

Factors: Case and Definiteness of argument2; 2-argument structures as fillers

nom-acc-dat, def-indef-indef vs. nom-acc-dat, def-def-indef vs. nom-dat-acc, def-indef-indef vs. nom-dat-acc, def-def-indef

Der Doktor wird einen/den
einem/dem Krankenpfleger einem/einem
einen/einen Rollstuhlfahrer zeigen
the_{nom} doctor will a/the_{acc/dat} (male) nurse a_{dat/acc} wheel chair-user point out to

Reading times (in ms)



Anova

Significant main effect of Case on Det3.
 Nothing else.

Discussion

Subjects don't expect a 3rd argument to follow after *nom-acc*. In accordance with the completion data. Not in accordance with the corpus data.

Corpus data

5 Syntactic frequency

Aggregated Negra 2 & Tiger Corpus

(syntactically annotated newspaper corpora, 80151 main and subclauses extracted) 4737 sentences with subject in the Vorfeld, main verb in final position, no pronouns

Results

nom-dat: 336 *nom-dat-acc*: 176
nom-acc: 4205 *nom-acc-dat*: 20

Discussion

2-argument structures are more frequent than 3-argument structures. As to 2-argument-structures, *nom-acc* is much more frequent than *nom-dat*. In 3-argument-structures, Datives tend to precede Accusatives. Kempen & Harbusch's (2003) data on sentences with all arguments in the midfield are supported by a greater database of sentences with the subject in the *Vorfeld*. But: The syntactic frequency data do not agree with the completion data.

6 Lexical frequency

Lexical (verb) frequency might account for the behavioral data.

Frequencies by subcategorization frames are extracted from the CELEX database.

Results

Summed verb frequencies (obligatory arguments, no word order information):
nom-acc: 481729 *nom-dat*: 220713 *nom-dat-acc, nom-acc-dat*: 90891
 Number of lemmas (obligatory arguments, no word order information):
nom-acc: 6336 *nom-dat*: 234 *nom-dat-acc, nom-acc-dat*: 662

Discussion

Subcategorization frame frequency as sum of the frequencies of the subcategorizing verbs cannot account for the completion data, as *nom-dat* is much more frequent than *nom-dat-acc*. But the number of lemmas reflects the behavioral data as *nom-acc* > (*nom-dat-acc, nom-acc-dat*) > *nom-dat*. Extent of choice seems more important than frequency.

7 Syntactic frequency reconsidered

Constraints on argument structures

Animacy was ignored in the syntactic frequency counts, but the completion and reading materials referred to animate entities only. Definiteness had no effect on completions and reading times, but it had one on the acceptability ratings reported by Featherston (2005).

Manual annotation of arguments

4737 sentences from syntactic frequency counts reported above animate vs. inanimate, definite vs. indefinite

Results

Number of sentences with DP1 and DP2 animate only:
nom-dat: 85 *nom-dat-acc*: 130 *nom-acc*: 452 *nom-acc-dat*: 0
 Definiteness does not interact with Case, Animacy or Argument Structure.

Discussion

Syntactic frequency analysis constraint to phrases with animate referents only finally reflects the reading and completion data. Definiteness has no effect on either corpus or behavioral data.

8 Determiners of syntactic frequency

Loglinear analyses

4737 sentences from the syntactic frequency counts, examined factors are: Animacy1 (1st argument), Animacy2 (2nd argument), Case (2nd argument), Argument Structure (2- vs. 3-)

Results: Weight of factors

Argument Structure*Case > (Animacy2*Case, Animacy2, Argument Structure) > Argument Structure*Animacy2

Discussion

Syntactic frequency is mostly determined by the interaction between Argument Structure (2- vs. 3-) and Case (2nd argument Dative vs. Accusative). However, the importance of Animacy for argument structure is restated. This can only partially be explained by the coincidence of Animacy and Dative Case: While 2/3 of the Datives in 3-argument structures refer to animate entities, this holds for only 1/2 of *nom-dat* structures. Thus, syntactic prevalence cannot be explained by syntactic features as Case alone, but semantic features like Animacy must be taken into account as well.

9 General discussion

Behavioral data

Argument structure expectations in sentences with the main verb in final position profit from a rich evaluation of argument-specific information as Case and Animacy. By contrast, Definiteness does not seem to play a role.

Corpus data

Corpus counts may serve as a predictor of subjects' argument structure preferences. But they have to take syntactic and semantic information into consideration.

Conclusions

One must rely on syntactically and semantically annotated corpora to predict subjects' behavior during sentence processing.

[CELEX] Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium.
 Featherston, S. (2005). *Experiment on word order in the Mittelfeld*. [http://www.sfb441.uni-tuebingen.de/~sam/db/wotan_exp.html]
 Friederici, A. D. & Frisch, S. (2000). Verb argument structure processing: The role of verb-specific and argument-specific information. *Journal of Memory and Language*, 43, 476-507.
 Kempen, G. & Harbusch, K. (2003). An artificial opposition between grammaticality and frequency: Comment on Bornkessel, Schlesewsky and Friederici (2002). *Cognition*, 90, 205-210.
 Kempen, G. & Harbusch, K. (2004). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann & C. Habel (Eds.), *Multidisciplinary approaches to language production* (pp. 173-181). Berlin: Mouton De Gruyter.
 Lapata, M., Keller, F., & Schulte im Walde, S. (2001). Verb Frame Frequency as a Predictor of Verb Bias. *Journal of Psycholinguistic Research*, 30, 419-435.
 [Negra2] http://www.coli.uni-sb.de/sfb378/negra-corpus/
 Rösler, F., Pechmann, T., Streh, J., Röder, B., & Hennigshausen, E. (1998). Parsing of sentences in a language with varying word order: Word-by-word variations of processing demands are revealed by event-related brain potentials. *Journal of Memory and Language*, 38, 150-176.
 [Tiger] http://www.ims.uni-stuttgart.de/projekte/TIGER/
 Weber, A. & Müller, K. (2004). Word order variation in German main clauses: A corpus analysis. *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva.