

# Availability of subcategorization frames: A matter of syntactic or lexical frequency?

Sandra Pappert<sup>1</sup>, Johannes Schließer<sup>1</sup>, Dirk P. Janssen<sup>2</sup>, & Thomas Pechmann<sup>1</sup>

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

Poster presented at Verb Workshop 2005, Saarbrücken  
Printed at the Computer Center of Leipzig University

## Overview

### 1 Introduction

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

- **Behavioral data:** Dative before Accusative is preferred (e.g., Rösler et al., 1998). But: Animacy is a confound!
- **Corpus data** on argument structures in the midfield (Kempen & Harbusch, 2003, 2004): *nom-acc* > *nom-dat* > *nom-dat-acc* > *nom-acc-dat* [ > means 'is more frequent than'] Animacy influences word order, too, but data on 3-argument-structures are sparse.

#### Hypotheses

- 2-argument-structures are preferred to 3-argument-structures.
- After *nom-acc*, this preference is more pronounced than after *nom-dat*.

### 2 This study

Does argument-specific information modulate subcategorization frame availability?  
Can behavioral data be predicted by syntactic or lexical frequency?

- **Materials**  
Word order: subject in the Vorfeld (topicalized), main verb in final position  
Arguments: unambiguous case marking, animate referents
- **Behavioral data**  
Sentence completion questionnaire as (comprehension and) production task  
Self-paced reading experiment as comprehension task
- **Corpus data**  
Syntactic frequency of argument structures from Negra2 and Tiger  
Lexical frequency for verbs with specific subcategorization frames from CELEX

## Behavioral data

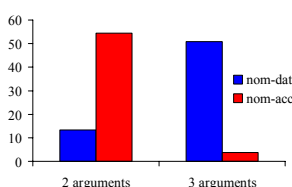
### 3 Sentence completion

#### Method

32 subjects, 32 sentence fragments, conditions *nom-dat* vs. *nom-acc*

Der Doktor wird *dem* Krankenpfleger ...  
*den*  
the<sub>nom</sub> doctor will the<sub>dat/acc</sub> (male) nurse ...

#### Completions (in %)



#### Discussion

As predicted: *nom-acc* > *nom-acc-dat*  
Not as predicted: *nom-dat-acc* > *nom-dat*  
We found no evidence for a correlation of syntactic frequency (cf. Kempen & Harbusch, 2003) and the availability of subcategorization frames.  
Will we find one in an online task like self-paced reading?

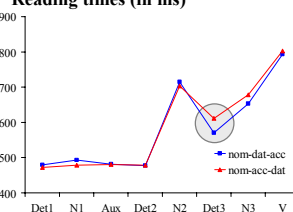
### 4 Self-paced reading

#### Method

32 subjects, 32 double object sentences, conditions *nom-dat-acc* and *nom-acc-dat*  
32 single object fillers to prevent subjects from predicting sentences' length

Der Doktor wird *dem* Krankenpfleger *den* Rollstuhlfahrer zeigen  
*den*  
the<sub>nom</sub> doctor will the<sub>dat/acc</sub> (male) nurse the<sub>acc/dat</sub> chair-bound point out to

#### Reading times (in ms)



#### Discussion

No difference on Determiner2  
Difference on Determiner3:  
*nom-acc-dat* > *nom-dat-acc*  
Subjects are more surprised to read a 3<sup>rd</sup> argument after *nom-acc* than after *nom-dat*.  
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-sentences are more frequent than 3-argument-sentences.  
As to 2-argument-sentences, *nom-acc* is much more frequent than *nom-dat*.  
In 3-argument-sentences, datives tend to precede accusatives.  
Kempen & Harbusch's (2003) data 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.

### 7 Syntactic frequency reconsidered

- **Semantic properties** were ignored in the syntactic frequency counts, but the completion materials consisted in animate DPs only.
- **Annotation of DPs** in the 4737 sentences from the syntactic frequency analysis: animate vs. inanimate, in case of a conflict of lexical (default) meaning and phrasal context counted as animate
- **Number of sentences with DP1 and DP2 animate only**  
*nom-dat*: 85      *nom-dat-acc*: 130  
*nom-acc*: 452      *nom-acc-dat*: 0
- **Discussion**  
Syntactic frequency analysis constraint to phrases with animate DPs only finally reflects the completion data: *nom-acc* > *nom-dat-acc* > *nom-dat* > *nom-acc-dat*.  
Semantic features of arguments seem to influence subcategorization frame availability.

### 6 Lexical frequency

• **Lexical (verb) frequency** might account for the behavioral data.  
=> Frequencies of subcategorization frames in CELEX database

#### Results

Summed verb frequencies (obligatory arguments, no constituent order information):  
*nom-acc*: 481729      *nom-dat*: 220713      *nom-dat-acc, nom-acc-dat*: 90891  
Number of lemmas (obligatory arguments, no constituent 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.

### 8 Determiners of syntactic frequency

- **Loglinear analyses** reveal the weight of factors for syntactic frequencies. Examined factors are: animacy of DP1, animacy of DP2, case of DP2, existence of DP3
- **Results**  
DP3exist\*caseDP2 > animDP2\*caseDP2, animDP2, DP3exist > DP3exist\*animDP2
- **Discussion**  
Syntactic frequency is mostly determined by an interaction of existence of DP3\*case of DP2.  
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 are animate, this holds for 1/2 of *nom-dat* structures. Thus, syntactic prevalence cannot be explained by syntactic features alone, but semantic features like animacy must be taken into account as well.

### 9 Conclusions

#### Behavioral data

Subcategorization frame availability in sentences with the main verb in final position profits from a rich evaluation of argument-specific information as case and animacy.

#### Corpus data

Only corpus counts that take syntactic and semantic information into consideration may account for behavioral data.

[CELEX] Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX Lexical Database* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium.  
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.  
[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/>