A quantitative corpus study of German word order variation

Kris Heylen[1]
QLVL – University of Leuven, Belgium
kris.heylen@arts.kuleuven.ac.be

1 Introduction

There is a growing awareness among theoretical linguists from different traditions that there is a problem of objectivity and reliability with traditional introspective data like individual grammaticality judgments or an individual researcher's assessments of examples. One of the options that linguists have pursued to get a firmer empirical basis for their research is the analysis of usage data as it is increasingly available in large corpora. However, usage data reflects a highly complex interplay of linguistic factors which cannot be analysed without the use of advanced statistical techniques. This paper presents a study into a specific type of word order variation in the Mittelfeld of the German clause as a case study of how a quantitative, statistically based corpus analysis can give a better insight into linguistic phenomena and what its limitations are.

1.1 Linguistic evidence and German word order variation

German clause topology is characterized by the so-called Klammer-construction. Two fixed positions, called Klammer and occupied by elements of the verbal group or by a complementizer, subdivide the clause into three main "fields". Of interest here is that the field between the two fixed positions, called the Mittelfeld, can contain multiple constituents and these constituents do not always occur in the same order. Especially the relative order of verb arguments co-occurring in the Mittelfed has been the subject of a lively debate within the German linguistics community, in which problems of linguistic evidence have played a major role. At the debate's climax in the mid 1980's, linguists mainly differed in opinion as to whether the word

[1] Kris Heylen is Research Assistant of the Fund for Scientific Research – Flanders , Belgium (FWO) and member of the QLVL-research group, more info: http://wwwling.arts.kuleuven.ac.be/qlvl/

order variation in the Mittelfeld was mainly determined by grammatical or pragmatic factors (see Reis 1987 for an overview). Both sides kept on coming up with contradictory examples whose validity was then put into question by the other side. The problem seemed to be that a great many factors were involved and each individual factor rarely had a categorical effect. This made unambiguous grammaticality judgements extremely difficult. Towards the beginning of the 1990's there was an increasing awareness that the main problem was indeed a methodological one and that the traditional introspective data was unreliable and could not cope with the phenomenon's complexity. As a consequence, a number of new empirical methods were tried out.

A first of these approaches used psycholinguistic experiments based on processing time differences (Pechmann et al. 1996, Poncin 2001), a second type of studies looked at corpus material (Primus 1994, Kurz 2000). A third and most recent approach uses a sophisticated version of grammaticality judgments, taken from multiple subjects and analyzed with advanced statistical techniques (Keller 2000, Featherston[2]). Yet, these approaches do have problems of their own. Both the psycholinguistic experiments and the corpus studies kept on struggling with the variation's multi-factor complexity: the psycholinguistic studies had to limit the number of factors because of their time-consuming data collection. The corpus studies could investigate multiple factors but lacked the statistical apparatus to deal with multifactorial phenomena. The third method of enhanced grammaticality judgments has enough data and the appropriate statistics to deal with multifactorial phenomena, but the heuristic status of grammaticality judgments themselves is not unproblematic. They allegedly reflect competence rather than performance, but this division itself is not uncontested in the linguistic community. The study presented in this paper is based on corpus material, i.e. usage-data. Improving on previous corpus studies, it will use multivariate statistics to analyze the interplay of multiple factors.

1.2 A specific type of word order variation.

This study focuses on a specific type of word order variation in the Mittelfeld, viz. the variation that occurs when both a nominally realized subject and a pronominally realized object are present in the Mittelfeld. In this case the pronominal object can either precede the nominal subject (ex. 1) or follow the nominal subject (ex. 2)[3].

(1) Ein paar Tage später nahm <ihn (OBJ)> <der SED-Chef der Uni (SUBJ)> beiseite
A few days later the university's SED-chief took him aside (NEGRA, 1618)

(2) Später, als <die Kommission (SUBJ)> <ihn (OBJ)> entlassen hat, sagt er,...
Later, when the commission has dismissed him, he says.... (NEGRA, 1665)

[2] ongoing, for more information: http://www.sfb441.uni-tuebingen.de/~sam/db/wotan.exp.html

[3] Examples from the NEGRA-corpus (see footnote 4) and referred to by corpus sentence number

Although the word order in example 1 is more common, both word orders seem to be freely interchangeable without any obvious difference in grammaticality or meaning. In this case, traditional heuristic methods like grammaticality judgments cannot discriminate between examples and therefore cannot detect the effect of relevant factors. Even the method of enhanced grammaticality judgments cannot detect a difference in acceptability between the two variants (Keller 2000: 108ff). The few other studies that discuss this type of variation (Lenerz 1994, Zifonun 1997:1511ff) admit that influencing factors are hard to identify. Although this variation seems even more elusive than other types, it is also a good starting point because it reduces the complexity of the variation by keeping one verb argument pronominal: pronouns vary less in length, given-new status and lexical diversity.

2 A quantitative corpus study

2.1 The corpus data

The study used the NEGRA corpus[4] (20602 sentences or 355096 tokens) which consists of morpho-syntactically annotated German newspaper text. The relevant observations, i.e. clauses with a nominal subject and a pronominal object in the Mittelfeld, were extracted semi-automatically and manually checked for precision and recall. There was a total number of 995 observations (distributed over 1015 sentences), which means the construction is rather common and present in 5% of the corpus' sentences. The observations were then annotated for the response variable word order (subject first vs. object first) and for a number of factors that are mentioned as relevant in the literature on word order variation and of which the following will be dealt with here: case of the pronominal object (accusative/dative), the subject's thematic role (agent/recipient/theme), length difference between subject and object (in number of syllables), given/new status (on a referent-accessibility scale[5] from 1 to 9), subject animacy (animate/inanimate), object pronoun type (personal/reflexive) and clause type (main vs. subordinate clause).

2.2 Statistical analyses

The effect of the above mentioned factors on word order was assessed through the use of different statistical techniques. Firstly, a univariate analysis of the distribution of the two word orders confirmed pronominal object before nominal subject as the default order (89%). Secondly, χ^2-tests showed that all factors had a very significant effect ($p<0.01$) on word order, except for the object's case ($p>0.9$). Next, Cochran-Mantel-Haenszel statistics allowed examining the effect on word order of one factor

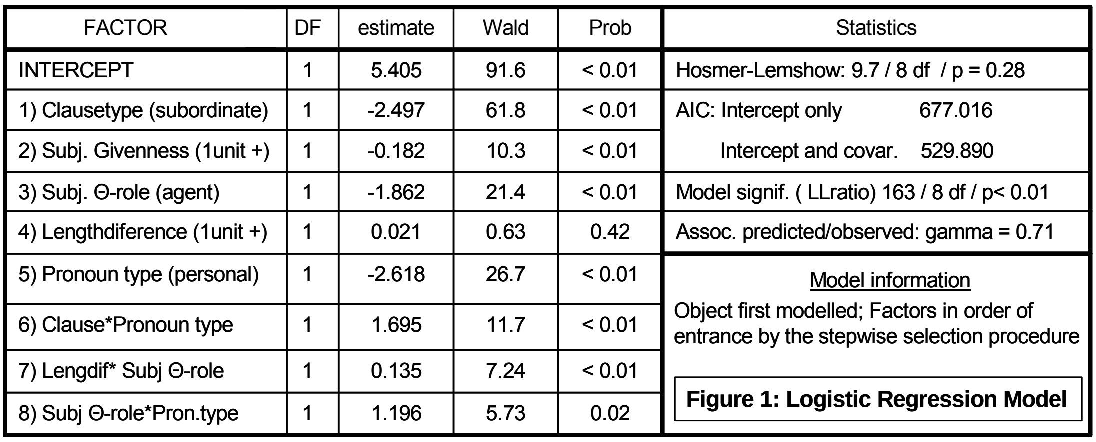
[4] compiled at the Univ. of Saarbrücken, more info at http://www.coli.uni-sb.de/sfb378/negra-corpus/

[5] adapted from Grondelaers 2000 and described in Heylen & Speelman 2003

while controlling for a second factor. Finally, more advanced multivariate statistical techniques like logistic regression (fig.1) and Classification and Regression Trees (CART, fig.2) allowed to examine the combined effect of multiple factors and their possible interactions, the relative importance of these factors and they allowed to assess how good the variation could be accounted for by reference to these factors. The logistic regression model shows the simultaneous effect of factors with negative estimates indicating an inhibitory effect on object first ordering. The classification tree shows the cumulative effect of factors in trying to separate object first and subject-first observations as much as possible in the leaves of the tree. Both techniques show that Clause type has the strongest effect (first entrance in the regression's stepwise procedure, first split in the tree). Furthermore, the classification tree tells us in which context there is most variation (cases with agentive, short subjects in subordinate clauses, second leave from the right), but also indicates that the factors analysed here, though significant, do not separate the data very well and that other factors will have to be considered in future studies.

3 Conclusion

Statistically based corpus studies like the one presented here can give insights into data that traditional types of linguistic evidence could not and certainly not with the same reliability. Yet, it is important to realize that these insights are not in themselves explanations of linguistic phenomena. Rather, they help to reveal underlying facts in the data on which linguistic explanations can be based. They help to uncover interesting correlations that would not have been apparent at first sight. As such, they provide the empirical basis that serves as an input and touchstone to theoretical linguistics. The findings of these corpus-based analyses help to focus theoretical research. However, the resulting theoretical hypotheses will eventually need further support from psycholinguistic and neurological experimental data, since the ultimate explanation for linguistic phenomena will have to be stated in terms of human cognitive abilities.



FACTOR	DF	estimate	Wald	Prob	Statistics
INTERCEPT	1	5.405	91.6	< 0.01	Hosmer-Lemshow: 9.7 / 8 df / p = 0.28
1) Clausetype (subordinate)	1	-2.497	61.8	< 0.01	AIC: Intercept only 677.016
2) Subj. Givenness (1unit +)	1	-0.182	10.3	< 0.01	Intercept and covar. 529.890
3) Subj. Θ-role (agent)	1	-1.862	21.4	< 0.01	Model signif. (LLratio) 163 / 8 df / p< 0.01
4) Lengthdiference (1unit +)	1	0.021	0.63	0.42	Assoc. predicted/observed: gamma = 0.71
5) Pronoun type (personal)	1	-2.618	26.7	< 0.01	Model information Object first modelled; Factors in order of entrance by the stepwise selection procedure
6) Clause*Pronoun type	1	1.695	11.7	< 0.01	
7) Lengdif* Subj Θ-role	1	0.135	7.24	< 0.01	
8) Subj Θ-role*Pron.type	1	1.196	5.73	0.02	

Figure 1: Logistic Regression Model

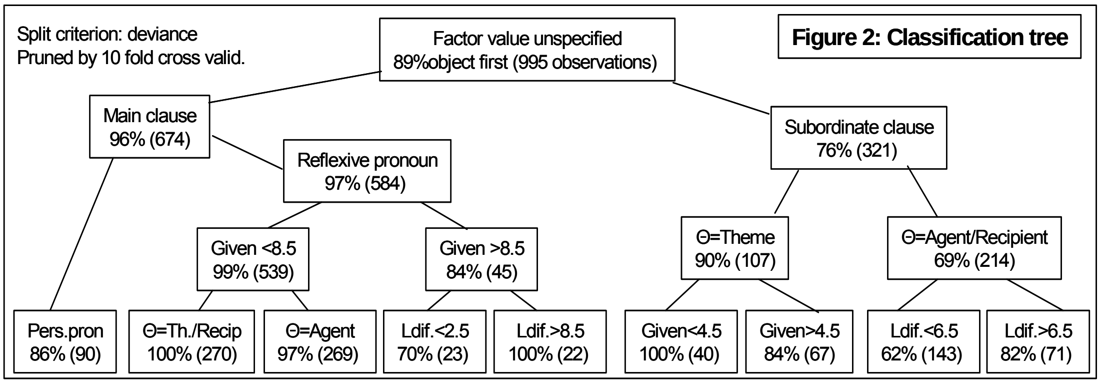


Figure 2: Classification tree

References

Heylen, K. and Speelman, D. (2003). A corpus-based analysis of word order variation: the order of verb arguments in the German middle field. In D. Archer et al., eds., *Proceedings of the Corpus Linguistics 2003 conference,* pp. 320-329.

Keller, F. (2000). *Gradience in Grammar. Experimental and Computational Aspects of Degrees of Grammaticality,* Unpubl. PhD thesis. University of Edinburgh.

Kurz, D. (2000). *Wortstellungsphänomene im Deutschen,* Unpubl. Master thesis, Universität Saarbrücken.

Lenerz, J. (1994). Pronomenprobleme. In B. Haftka, ed., *Was determiniert Wortstellungsvariation? Studien zu einem Interaktionsfeld von Grammatik, Pragmatik und Sprachtypologie,* pp 161-173.

Pechmann, T. et al. (1996). Wortstellung im deutschen Mittelfeld. Linguistische Theorie und psycholinguistische Evidenz. In C. Habel et al. , eds., *Perspektiven der kognitive Linguistik. Modelle und Methoden,* pp 258-299.

Poncin, K. (2001). Präferierte Satzgliedfolge im Deutschen: Modell und experimentelle Evaluation. *Linguistische Berichte*, 186:175-203.

Primus, B. (1994). Grammatik und Performanz: Faktoren der Wortstellungsvariation im Mittelfeld. *Sprache und Pragmatik*, 32: 39-86.

Reis, M. (1987). Die Stellung der Verbargumente im Deutschen. Stilübungen zum Grammatik:Pragmatik-Verhältnis. In I. Rosengren, ed., *Sprache und Pragmatik: Lunder Symposium 1986*, pp 139-177.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). *Grammatik der deutschen Sprache*, de Gruyter. Berlin / New York.