A corpus-based analysis of word order variation: The order of verb arguments in the German middle field

Kris Heylen[1] and Dirk Speelman
University of Leuven

Abstract

A perennial problem in German syntax is the order of verb arguments in the Mittelfeld. The Mittelfeld is the section of the clause between the two parts of the discontinuous verbal group. In it, all verb arguments can be realized simultaneously, however, not always in the same order. There is a long-standing debate about a number of factors that possibly govern this variation, yet, what their actual influence is, remains unclear. This study takes a quantitative corpus-based approach to the problem and looks specifically at a type of variation that has scarcely been dealt with up to now, viz. the relative order of a pronominal object and a nominal subject in the Mittelfeld. Clauses that show this kind of variation have been collected from the NEGRA-corpus consisting of German newspaper material and they have been annotated for 5 factors frequently mentioned in the literature: grammatical function of the arguments, given/new status and animacy of the arguments' referents, difference in length between the arguments, and their occurrence in either a main or a subordinate clause. The effect of these factors has been statistically checked and modelled in a logistic regression model. The results of the statistical analysis show an effect of all factors except for grammatical. The effect of main versus subordinate clause is especially strong, contradicting earlier hypotheses that this factor is only an epiphenomenon of length difference

1 Introduction

German clause topology is characterized by the so-called *Klammer* construction. Within the clause, verbal elements have two fixed positions that seem to hold the clause together like two braces (German: Klammer). Using this representation scheme, other constituents can be positioned relative to the two Klammers: the site before the first Klammer is called Vorfeld (front field), the one between the two klammers is referred to as the Mittelfeld (middle field), and the position after the second Klammer is known as the Nachfeld (back field). Table 1 shows that in the main clause, the first Klammer is occupied by a finite verb, in this case *hat*, whereas the second Klammer is filled by infinite verbal elements, here the past participle *geschenkt*. The Vorfeld has room for one constituent (*Gestern*) and the remaining constituents normally occupy the Mittelfeld. Very long constituents or, like in this case, subordinate clauses are placed in the Nachfeld. The subordinate clause has a somewhat different structure because all of the verbal elements, both finite and infinite, now occupy the second Klammer, whereas the first Klammer is filled by the complementizer (*dass*) or is left empty. The front field is empty too so that all of the clause's constituents land in the Mittelfeld.

		Vorfeld	*1st Klammer*	*Mittelfeld*	*2nd Klammer*	*Nachfeld*
main clause		Gestern	Hat	der Vater dem Sohn einen Ball	geschenkt,	weil er Geburtstag hatte
	Yesterday, the father has given the son a ball because it was his birthday					
subordinate clause	(Er sagte,)		dass	der Vater dem Sohn gestern einen Ball	geschenkt hat,	weil er Geburtstag hatte
	(He said) that the father has given the son a ball yesterday because it was his birthday					

Table 1

The order of constituents in the Mittelfeld is not fixed. Especially the order of the verb arguments is subject to considerable variation. Nominative subject, dative object and accusative object can occur simultaneously in the Mittelfeld. Although the default order is the one shown in Table 1 with

[1] Kris Heylen is Research Assistant of the Fund for Scientific Research - Flanders

nominative < dative < accusative, other sequences are possible as well. The question which principles determine this word order has led to an animated debate within the German linguistics community, starting with Lenerz's (1977) seminal study and reaching a climax in the mid eighties (see Reis 1987 for an overview). In this discussion functionally oriented linguists opposed advocates of an autonomous syntax. The former stressed the importance of pragmatic factors, whereas the latter insisted on purely grammatical principles. As both camps kept on coming up with examples that contradicted previous examples from the opposite side, the debate was never settled. Since then, a number of attempts have been made to examine the variation more empirically. On the one hand there have been psycholinguistic studies looking at the influence on processing time of reordering constituents with different grammatical functions (Pechmann et al. 1996, Poncin 2001). On the other hand there have been small-scale corpus studies that following Hawkins (1994) have looked at syntactic complexity, i.e. length, as a possible fundamental principle regulating word order (Primus 1994, Kurz 2000). Yet, despite of all this research the factors determining the word order variation in the Mittelfeld remain unclear.

The most important reason for this lack of insight is certainly the variation's complexity. So many different factors are involved that Reis (1987: 139) speaks about a "jungle of factors" and Eisenberg (1994: 419) refers to a "welter of criteria". However, the research up to now has had shortcomings of its own that prevented an adequate description of this complexity. First of all, the different studies used disparate methods like grammaticality judgements, psycholinguistic experiments or corpus studies, which make the results very difficult to compare. Apart from that, the different methods had weaknesses of their own. The bulk of the studies from the eighties but also some more recent ones use introspection and grammaticality judgements of constructed examples as research data. Apart from the precariousness of using constructed examples, it is also problematic that the factors rarely have a categorical effect that leads to clear ungrammaticality. This often makes the assessment of a factor's effect vere subjective. The psycholinguistic experiments on the other hand can rely on clear factual evidence. However, their measurements in terms of latencies cannot always be easily linked to the different linguistic variables that play a role in the complex variation. This caused Pechmann et al. (1996) and Poncin (2001) to restrict themselves to only one factor, which they could control and gather enough data for. Finally, the corpus studies carried out by Primus (1994) and Kunz (2000) did investigate 2 or 3 factors but their datasets were rather small and most of all, they lacked the statistic apparatus to analyse multivariate data. The study presented here tries to learn from the weaknesses of previous studies by opting for a quantititave corpus study, based on the statistical analysis of actual usage data.

The current study will focus on a specific type of variation in the Mittelfeld that up to now has been scarcely taken into account. When the Mittelfeld of a clause contains both a subject realized as a full NP and an object realized as a personal pronoun, the pronominal object most frequently precedes the nominal subject. In a substantial number of cases however, the pronominal object follows the nominal subject. Moreover, the two orders seem freely interchangeable without any obvious meaning difference between them. In the literature pronominality is often mentioned without any further consideration as a factor that stimulates pre-posing of constituents. Somewhat more in-depth studies are Lenerz (1994) who looks at this type of variation from a Government & Binding perspective, only to find that this must be some form of scrambling, and Shannon (2000), who uses corpus data, but without further analysis for the synchronic variation, since he is mainly interested in the diachronic analysis. In one respect, the word order variation of nominal subjects and pronominal objects is even harder to get at because it does not show the differences of meaning or grammaticality, that can still be observed up to a certain degree with full-NP subjects and objects. That is probably the reason why this type of variation is scarcely studied and often dismissed as free variation. Yet in another respect, controlling for pronominality and restricting the variation to pronominal objects and nominal subjects reduces the complexity of the problem considerably. Pronouns vary less in length and given/new status, two factors that have been central in the debate. This means we can concentrate on the subject as the most prominent verb argument. The purpose of this study is therefore twofold: On the one hand it tries to gain some insight into a less studied type of variation, but on the other hand it uses this type of variation as a starting point to study the word order variation in the Mittelfeld in general. By initially reducing the complexity and making the problem more amenable to examination, we hope to get an easier grip on the more complex types of variation afterwards.

In practice, this study will look at 5 factors that are frequently mentioned as having an influence on the order of verb arguments in the Mittelfeld: the grammatical function of the arguments, their length, their occurrence in a main versus a subordinate clause (clause type), their given/new status, and the animacy of their referents. We will examine what the actual effect of these factors on the word order is; what

their combined effect is; whether there is any interaction between factors; whether some factors are more important than others; and finally, how much of the variation is explained by these factors.

We will start off with a short description of the corpus material. Next, we discuss the individual effect of the factors and some two-factor interactions. Then, we use a logistic regression model to look at the factors' combined effect. Finally, we conclude with a round-up of the results and give some suggestions for future research.

2 Corpus material

In the choice of a corpus two issues play an important role: the representativeness of the corpus and its accessibility. Representativeness relates both to what kind of language use has been included in the corpus and to the number of instances of the construction under investigation. That number should be large enough to base any reliable conclusions on. Accessibility on the other hand does not only pertain to the obvious fact that (machine-readable) corpora should be made available to the academic community. It also relates to the corpus's mark-up and annotation. An extensive annotation makes it easier to trace more complex syntactic structures. After comparing a number of electronic German corpora, the NEGRA-corpus seemed to offer the best compromise for our study between these two concerns. The NEGRA-corpus[2] was compiled at the university of Saarbrücken (Germany) and is made up of newspaper material from a local newspaper called *Frankfurter Rundschau*, year 1991-1992. This means that, strictly speaking, any conclusions we reach can only be valid for this type of language use. Yet, newspapers do function actively in a language community so that they might be considered appropriate to base a first study on. The NEGRA-corpus contains some 355,015 tokens or 20,602 sentences. That is not an awful lot, but the construction we are looking at will appear to be common enough to yield a sufficient number of instances for analysis. The main advantage of the NEGRA-corpus over the other corpora is its annotation. It has been annotated and manually checked for Part-of-speech values and syntactic structure. This allows for an almost automatic detection of the Klammer construction, which is central to German word order description.

A first step in preparing the corpus material for further study, was to extract the observations containing the relevant construction. To do so, PERL-scripts were written that exploited the NEGRA corpus's morpho-syntactic annotation. Relevant observations were defined as all clauses that contain a nominal subject and a pronominal object in their Mittelfeld. A nominal subject had to be either a full NP, containing a common or a proper noun, or a demonstrative or indefinite, which are known from the literature (Zifonun 1997: 1510) to have the same topological behaviour as nouns. Pronominal objects had to be either accusative or dative objects. Clauses with both an accusative and dative object were so rare that they could not be taken into account for statistic analysis. Furthermore, pronominal objects had to be realized as personal pronouns, including reflexives. As just said, demonstrative and indefinite pronouns have a noun-like topological behaviour. The output of the scripts was manually checked for spurious hits, but they were rare (4%). Part of the corpus was controlled manually to see whether the scripts had missed relevant observations, but that was not the case. In the end, 995 observations were retained, which means that almost 5% of all sentences in the corpus exhibited the construction under investigation.

In a second step, we annotated each of the observations by recording how each of the 5 factors was realized for that observation. An annotation scheme was set up with for each factor the possible values it could take. The actual annotation was done semi-automatically using the *Abundantia Verborum* annotation tool (Speelman 1997). The annotation procedure took the form of an annotation loop: When an observation did not fit into the initial annotation scheme, the scheme was revised to accommodate for that observation. Then, all previously annotated observations were checked to see whether they were still in compliance with the new annotation scheme. This reiteration went on until all observations were annotated adequately. Finally, the outcome of the annotation was turned into a data matrix with one line per observation stating the values observed for each factor.

3 Statistical analysis of individual factors

The ultimate purpose of this study is to determine the influence of the 5 factors, individually and combined, on the word order variation of pronominal objects and nominal subjects in the Mittelfeld. This is done through statistical analysis. Statistical analysis offers two main advantages: it generalizes over a lot of examples and it tells you how sure you can be of your results. Generalizing over examples

[2] More information about the corpus is available under http://www.coli.uni-sb.de/sfb378/negra-corpus/

is important to detect effects that are not apparent at first sight. In the case of word order variation in the Mittelfeld, none of the factors has a categorical effect leading to a clear difference in meaning or grammaticality. The influence these factors have is a matter of tendency, which is very hard to get at with the traditional heuristic methods like grammaticality judgements used in many of the earlier studies. By counting and generalizing over instances, these tendencies do become clear. The other advantage of proper statistical analysis is that it can verify the significance of the results obtained from a sample for the phenomenon in general. In this study, we will check whether the factor influences we observe in the corpus can safely be generalized for this type of language use.

The factors we study are so-called categorical variables. Their values are not measured on a continuous numerical scale but are discrete categories. Categorical data analysis requires specific statistical techniques. The statistics[3] used here can be split up in three categories. In this section we will first carry out a *univariate analysis*, looking at one isolated variable's distribution. This will be the variation in word order itself, the variable whose behaviour we are ultimately interested in, also called the response variable. Confidence interval will be calculated for its distribution. The remainder of this section is reserved for *bi-* and *trivariate analysis*. Here we investigate each time the relation between the response variable and one or two of the factors that we suspect have an influence on the response variable. They are referred to as the explanatory variables. Whether there is association between an explanatory variable and the response variable will be tested with the Pearson chi-square statistic. How strong that association is will be expressed by the odds ratio. For the ordinal variables length and given/new status, we will test for linear association using the Mantel-Haenszel chi-square statistic. Strength and direction of this linear association is conveyed by the gamma coefficient. In section 4 then, we will look at a *Multivariate analysis*. There we will try to examine what the combined effect of the explanatory variables is on the response variable, using a logistic regression model[4].

3.1 Response variable: word order

Let us first look at the response variable, viz. the variation in word order of nominal subjects and pronominal objects in the Mittelfeld of the German clause. The response variable can take two values: either the pronominal object precedes the nominal subject (object first as in1) or the pronominal object follows the nominal subject (subject first as in 2).

(1) Ein paar Tage später nahm <ihn (OBJ)> <der SED-Chef der Uni (SUBJ)> beiseite (NEGRA, 1618)[5]
A few days later the university's SED-chief took him aside

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

If we look at the distribution between object first and subject first as presented in Table 2, we see that object first is much more common. This confirms the status of object first as the default word order, which is often mentioned in the literature. To check how representative this distribution in the NEGRA-corpus is for newspaper language in general, we can look at the 95% confidence interval. For the probability of having object first this interval lies between 87.1% and 91.0%[6]. The confidence interval tells us with 95% certainty that the share of object first cases in this type of newspaper language in general will lie between 87.1% and 91.0%, which again confirms object first as the default order.

Object first	*Subject first*
889/995 (89,0%)	106/995 (11,0%)

Table 2

Yet, the distribution is much more askew then previous exploratory studies suggested. Hoberg's (1981) descriptive figures for her mixed corpus show a ratio of 75% to 25% based on 358 observations. A chi-square test on a cross tabulation of both datasets shows that this difference is very significant ($p<0.01$). Shannon's (2000) study of literary texts has a 62% to 38% ratio for object first versus subject first based on 786 observations and that difference is very significant too($p<0.01$). Moreover, the difference between Shannon's and Hoberg's corpus distribution is significant as well ($p<00.1$). It is not immediately clear what causes these differences in distribution. However, it is notable that Shannon's

[3] For an overview of the statistics for categorical data analysis used here, see Agresti (1996)

[4] For the analyses, the statistical software pack, SAS version 8 was used.

[5] Examples from the NEGRA corpus are referred to by "NEGRA" followed by the number of the sentence they appear in.

[6] For the calculation of the confidence interval for extreme proportions, see Agresti (1996: 11)

corpus mainly contains literary material, Hoberg's is a mix of all kind of written material, and ours is newspaper material. Therefore, these differences might be due to an influence of text sort.

3.2 Grammatical function

Grammatical function refers to the case role assigned to an argument, which is (mostly) morphologically marked in German. Because case has always been a central feature of syntactic descriptions, it is probably the oldest factor to be acknowledged as having an influence on the order of verb arguments in the Mittelfeld. Indeed, it is the factor used in grammars to tell the verb arguments apart and to assign ordering preferences to them. In this study, we use case as a basic descriptor as well. We defined the relevant construction as the one having a nominal, nominative subject and a pronominal accusative or dative object. Because the presence of a nominative subject is held constant, the values of the factor grammatical function pertain to the case of the pronominal object: this is either an accusative object or a dative object. Looking at the contingency table in Table 3, it is clear that accusative objects are more frequent than dative objects. This is not a surprise as there are many more transitive verbs than ditransitive ones. What is surprising though, is that the case of the object does not seem to have any influence at all on the distribution of the word order (chi-square, p=0.94). For both dative and accusative case object first has a proportion of about 89%. Although the literature (Lenerz 1977, Zifonun 1997) suggests for nominal objects, that the dative object has a greater tendency to precede the subject than the accusative object, this is not borne out for pronominal objects here. In general, grammatical function is seen as the most important factor of all for word order (Reis 1987, Primus 1994). Again, this does not seem to hold for the order of nominal subjects and pronominal objects.

	Object first	*Subject first*	*/total*
Accusative	724/810 (89.4%)	86/810 (10.6%)	810/995 (81,4%)
Dative	165/185 (89.2%)	20/185 (10.8%)	185/995 (18,6%)

Table 3

3.3 Length

Length refers to the difference in length between the pronominal object and the nominal subject as measured in syllables[7]. Difference of length was already mentioned by Behaghel (1932) as his "law of growing constituents" saying that longer constituents follow shorter ones. More recently, Hawkins (1994) stated in his EIC-theory that length difference is the main principle governing word order universally, and also in German. More specifically the theory[8] states that word order is determined by the attempt to minimize the time a listener needs to determine how many constituents a clause has. This can be done by putting longer constituents further behind in the clause A first word of a constituent, e.g. a preposition, is often enough to determine the nature of constituent (e.g. a prepositional phrase). If shorter constituents were to follow longer ones, the listener would have to go through all the words of that long constituent before he or she could determine the presence of the shorter constituent at the end. By putting the shorter ones first, a listener already knows about all the other constituents by the time he reaches the first word of the last constituent and from that word alone he can now determine the total number of constituents without having to go through all the words of the long constituent.

In Table 4 the length difference has been split up into 6 categories starting from no length difference and going up to larger differences in number of syllables. As pronouns in German are only 1 or 2 syllables long, increasing length difference is caused by longer subject NP's. Table 4 bears out the expected influence of length difference: The larger the difference between the pronominal object and the nominal subject gets, the more often the object precedes the subject. Statistically we can check this directional relation using the Mantel-Haenszel chi-square test, which confirms that there is linear association ($p<0.01$). The strength and direction of this association is given by the gamma coefficient, which has a scale from –1 to 1 with –1 being perfect negative association, 0 no association and 1 perfect positive association. Here a gamma coefficient of –0.4 (ASE=0.065) indicates moderate negative association, which means that higher length differences tend to correspond relatively more with object first and smaller length differences tend to correspond relatively more with subject first.

[7] Number of syllables was chosen over number of words as a measure because German has very long compound nouns, to which a number of words measure is not sensitive.

[8] The theory is presented somewhat simplified here. For an in-depth treatment see Hawkins (1994).

	Object first	*Subject first*
0	105/137 (76.6%)	32/137 (23.4%)
1-3	194/225 (86.2%)	31/225 (13.8%)
4-6	212/233 (91.0%)	21/233 (9.0%)
7-10	153/163 (93.9%)	10/163 (6.1%)
11-20	155/164 (94.5%)	9/164 (5.5%)
20-65	70/73 (95.9%)	3/73 (4.1%)

Table 4

3.4 Clause type: main versus subordinate clause

The factor of clause type, i.e. the occurence of the construction in a main versus a subordinate clause, has also been put forward in the context of Hawkins' (1994) EIC-theory explained above. According to Hawkins, the topological characteristics of German subordinate clauses diminishes the effect length difference has on word order. As explained in the introduction, the position of the finite verb in German is different in main clauses from subordinate clauses. In main clauses the finite verb is in second position, whereas in subordinate clauses, the finite verb is always near the end of the clause. In Hawkins's terms, this means a listener in subordinate clauses always has to wait until the end to be able to ascertain the total number of constituents, no matter whether the speaker puts longer constituents before shorter ones or not. That is why, according to Hawkins, the speaker will put relatively more shorter constituents before longer ones in main clauses than in subordinate clauses. This results in more object first cases in main clauses than in subordinate ones. Note that in this hypothesis the effect of clause type is subsidiary to length: clause type has only an influence because of length's initial effect.

	Object first	*Subject first*
main clause	646/674 (95.9%)	28/674 (4.1%)
subordinate clause	243/321 (75.7%)	78/321 (24.3%)

Table 5

If we look at the results in Table 5, we can indeed see that object first is significantly more common in main clauses than in subordinate clauses (chi-square $p<0.01$). An odds ratio of 7.4 (with confidence interval 4.7 to 11.7) tells us that the odds of object first versus subject first are 7.4 times higher in main clauses than in subordinate clauses, which indicates a very strong association. Considering the fact that the pronominal objects are generally shorter than the subjects, this domination of object first in main clauses would seem to confirm Hawkins' hypothesis. If we perform the analysis for length we did in the previous section again, but now separately for main clauses and subordinate clauses, we see that length has a significant influence in both clause types (chi-square $p<0.01$), but that, as Hawkins predicts, the effect is much stronger in main clauses (gamma = -0.57) than in subordinate clauses (gamma = -0.37). Yet, there are reasons to doubt Hawkins' assumption that clause type has only a subsidiary influence on the word order, i.e. dependent on length. For length we only measured a moderate association with word order (gamma = -0.4), whereas for clause type the association with word order was very strong (odds ratio of 7.4). It would be counterintuitive for a subsidiary effect to be stronger than the one it depends on. One can argue that the two measures of association are difficult to compare, but there is more. Table 6 shows the cross tabulation of word order and clause type for observations where there is no difference in length between the object and the subject. Here too, there is a strong effect of object first being more common in main clauses than in subordinate clauses (chi square $p<0.01$, odds ratio = 4.6 with confidence interval of 2.0 to 10.6). Yet, length difference cannot come into play here as the underlying explanation. This indicates that clause type is not just a subsidiary effect to length, but has a significant influence of its own. We will present more evidence for this independent effect of clause type in section 4 discussing the logistic regression model.

	Object first	*Subject first*
main clause	77/89 (86.5%)	12/89 (13.5%)
subordinate clause	28/48 (58.3%)	20/48 (41.7%)

Length difference=0 Table 6

3.5 Given/new status

Given/new status was initially introduced as a factor influencing word order by linguists of the Prague School in the 1930's. Since then, it has been commonly assumed that constituents referring to contextually known information precede constituents referring to new information. Lenerz (1977) considers given/new status also for German as central to the word order. However, it is not always clear what counts as given information and what as new. To make the given/new distinction operational, we have used the taxonomy designed by Grondelaers (2000), which estimates the givenness of referents by how mentally accessible they are to the listener. Table 7 demonstrates the scale of the given/new taxonomy ranging from an entity that has to be newly created in the mental representation of the listener (1) to an entity immediately accessible in context (8). Because referents of pronouns are almost always accessible in the near context, they were not taken into account. This allows us to concentrate on the given/new status of the subject's referent. Table 8 shows the distribution of word order over the different degrees of the given/new taxonomy. The percentage of subject first cases is not monotonic increasing as the givenness of the subject referent rises. The Mantel-Haenszel chi-square statistic indicates however, that there is a significant linear association ($p<0.01$). The gamma-coefficient of 0.28 (ASE 0.067) points at a moderate, positive linear association. This means that the more given a subject referent is, the more likely it is for the subject to precede the pronominal object. This confirms the general view that given information tends to precede new information.

Given /new taxonomy	
1	to be created
2	to be created but constrained by context
3	to be created, constrained by an anchor entity in context
4	accessible through encyclopaedic knowledge
5	inferable from anchor entity in context
6	accessible in wider linguistic context
7	inferable from near linguistic context
8	accessible in near context

Table 7

	Object first	*Subject first*
1	162/169 (95.9%)	7/169 (4.1%)
2	48/55 (87.3%)	7/55 (12.7%)
3	24/24 (100.0%)	0/24 (0.0%)
4	103/117 (88.0%)	14 (12.0%)
5	101/106 (95.3%)	5/106 (4.7%)
6	255/291 (87.6%)	36/291 (12.4%)
7	140/159 (88.1%)	19/159 (11.9%)
8	56/74 (75.7%)	18/74 (24.3%)

Table 8

3.6 Animacy

Animacy is the central principle determining word order within the Mittelfeld according to the functionally inspired reference grammar edited by Zifonun et al. (1997). Constituents with animate referents are said to have a tendency to stand at the front of the middle field. Because the reflexives among the pronominal objects "copy" the animacy of their subjects, results for the object's animacy would be somewhat misleading. Therefore, we concentrate on the animacy of the subject. Table 9 presents the distribution of word order for animate subjects versus inanimate ones. The results confirm that there is a significant association between animacy and word order (chi-square $p<0.01$). The odds ratio of 0.43 (confidence interval 0.27 to 0.70) tells us that the odds of object first versus subject first

amongst animate subjects is only half of what the odds are amongst inanimate subjects. In other words, animate subjects do get relatively more fronted than inanimate subjects.

	Object first	*Subject first*
animate subject	532/614 86.6%	82/614 13.4%
inanimate subject	357/381 93.7%	24/381 6.3%

Table 9

4 Multivariate analysis: logistic regression

After having verified the influence of each of the 5 factors separately, we will now look at their combined effect. This means that we will try to determine which factor is most important, what kind of interactions between factors show up and how good our model based on 5 factors is. This kind of multivariate analysis can be done using a logistic regression model, in the sociolinguistic literature also known as VARBRUL. This technique models the probability for one of the values of the response variable as a weighted linear function of the explanatory variables as in the formula:

logit(probality of response value1) = constant + weight1*factor1 + weight2*factor2 + weight3*factor3

Because probabilities can only lie between 0 and 1 and the function can take in principle any value, the logit (the odds ratio's natural log) of the probability is modelled instead. For our study, the logistic regression model will model the probability of having object first using the 5 factors discussed above as explanatory variables

We want the model to use only those explanatory variables that really explain some of the variation in the data. Statistical software can select these relevant explanatory variables using a stepwise selection procedure. The software will successively add more explanatory variables to the model by checking whether the model improves significantly compared to a model with only a constant called the intercept.. The software adds the variables that explain most of the variation first and stops adding variables when the model cannot be improved significantly any more..

Summary of Stepwise Selection				
Step	*Effect Entered*	*DF*	*Score Chi-Square*	*p-value*
1	Clause type	1	92.7053	<.0001
2	Given/new	1	17.0988	<.0001
3	Animacy	1	13.7755	0.0002
4	Length	1	10.5681	0.0012
5	Length*Animacy	1	24.9775	<.0001

Table 10

Table 10 shows the outcome for the stepwise selection procedure for all of the 5 factors and all of their two-by-two interactions. The final model contains 4 explanatory variables and 1 interaction term. The chi-square statistic shows for each step that a significant improvement was made. The model-of-fit statistics indicate that the model fits the data reasonably well (Pearson p=0.12, Hosmer & Lemeshow p=0.74). It appears that clause type has the biggest effect on the choice of word order, followed by the given/new distinction, animacy and length difference. As we could already expect from the bivariate analysis above, grammatical function did not enter the model. Apparently it has no significant influence on the variation. The stepwise procedure also confirms what we had noted in section 3.4, viz. that clause type seems to be far more important than length difference. This gives an additional indication that clause type is not a subsidiary effect of length difference but that it has an effect of its own. Even more so, it has the most important effect of all. Moreover, the interaction between length difference and clause type did not make it into the model, which would have been expected, had clause type been a subsidiary effect. Somewhat surprisingly, the interaction between length difference and animacy did enter the model. Apparently, the effect of length is bigger at a certain level of animacy or vice versa, animacy has a bigger effect for certain length differences. The nature of this effect will become clearer looking at the next table.

Table 11 shows the estimates for the weights each explanatory variable gets in the final model for predicting the probability of object first. The intercept represents the default result value. "Value" gives the value of the factor that was taken into account. The estimates indicate how strong the effect of a factor having a specific value is. Negative estimates point at a decrease in probability for object first, whereas positive estimates are a sign of increasing probability. The Wald chi-square and its p value designate whether a factor's effect is significant ($p<0.01$). The odds ratios give the effect the modelled value of a factor has on the odds of choosing object first over subject first. From these results we learn that the default value (the intercept) favours object first. Subordinate clause, given subjects and animate subjects disfavour object first, whereas a bigger length difference leads to more object first cases. Now the meaning of the interaction between length and animacy becomes clearer as well: bigger length differences combined with an animate subject lead to more object first cases.

Parameter	*Value*	*DF*	*Estimate*	*SE*	*Wald chi-square*	*p*	*Odds ratio*
Intercept		1	2.8903	0.4174	47.9386	<.0001	
Clause type	subordinate	1	-1.0597	0.1233	73.8087	<.0001	0.347
Given new	1 unit increase	1	-0.1682	0.0559	9.0487	0.0026	0.845
Animacy	animate	1	-0.9974	0.1812	30.3050	<.0001	0.369
Length	1 unit increase	1	0.1014	0.0253	16.0799	<.0001	1.107
*Length*animacy*	animate	1	0.1096	0.0252	18.8982	<.0001	1.116

Table 11

Finally, we come to the question, how good our model is. One way to assess the model is by looking at the explained variation. The Akaike Information Criterion gives a first measure for the unexplained variation in the model with only the intercept and a second measure for the unexplained variation in the model with the variables. The more measure decreases, the better. For our model we go from 677 to 551, which is a significant decrease. A second way of assessing the model is by looking at its predictive power. To measure this, we let the model predict word order in the original dataset and we look at how many cases the model predicted correctly. This we compare with the number of correct predictions an uninformed model with only a default parameter (the intercept) can make. The bigger the difference between these numbers, the better our model. The uninformed model predicted 89.1% of the cases correctly. The model with factors predicted 89.9% of the cases correctly at a cut-off point of 0.4. This tiny increase of 0.8% is of course somewhat disappointing, but not unexpected. The distribution of the response variable is so much askew (89% object first versus 11% subject first) that there is not much to it to have a good score on the prediction. You simply have to choose the most frequent value all the time and you get 89% correct. For the model with variables, it is very difficult to do much better. For a better assessment of the model we will have to use more data and other statistical techniques that are better suited for skewed data.

5 Conclusion

In this study we have looked at a long-standing problem in German syntax, viz. the word order variation of verb arguments in the Mittelfeld. We have analysed some of the problems previous studies encountered in studying this variation. We have proposed ways in which to deal with this very complex variation. On the one hand, we have chosen to restrict ourselves to one specific type of word order variation in the Mittelfeld, viz. the variation in the relative order of a nominal subject and a pronominal object. In this type of variation the complexity is somewhat reduced and therefore it provides a good starting point to study word order variation in the Mittelfeld in general. On the other hand we have argued that the influence different factors have is so subtle that a quantitative analysis on a firm empirical basis is called for, using extensive corpus data and the appropriate statistical apparatus. Next, we went on to present as a case study a first corpus analysis dealing with the word order variation of pronominal objects and nominal subjects in the Mittelfeld. First, we have been able to confirm that in general pronominal objects precede nominal subjects. Then we looked at 5 factors that according to the literature have an influence on the variation. Analysing their individual effects, we were able to show, that grammatical function is not as important as previously thought. Furthermore, length difference and clause type were both shown to have an influence, but clause type is not an epiphenomenon of length

difference, as previously thought. The given/new distinction and animacy did have the effect described in the literature. In the multivariate analysis we were able to show using a logistic regression model that clause type is the most important of the 5 factors in determining the relative order of nominal subjects and pronominal objects. Additional confirmation was given to contradict the hypothesis that clause type is a subsidiary factor to length difference. We have shown that our model does reduce the amount of unexplained variation. Predictive power however, is difficult to assess due to the skewness of the response variable

Some results from the analysis offered interesting suggestions for future research. We noted that in comparison with studies of literary text material the word order in our study of newspaper material is significantly more biased towards the default order. This might indicate an effect of text sort that has to be further looked into. The insignificance of grammatical function for the variation was very surprising as it is the one factor about which there is general agreement that it does matter. Further examination is needed to determine why its effect is not apparent here. Now we have shown that clause type has a strong effect, independent of length difference, the next step is to look for an alternative motivation behind it. Finally, an interesting interaction between length difference and animacy appeared in the model. What shape this interaction takes precisely, has yet to be determined.

References

Agresti A 1996 *An Introduction to Categorical Data Analysis.* New York, John Wiley & Sons.

Behaghel O 1932 *Deutsche Syntax: eine geschichtliche Darstellung. 4: Wortstellung, Periodenbau*. Heidelberg, Winter.

Eisenberg P 1994 *Grundriß der deutschen Grammatik. 3., überarbeitete Auflage.* Stuttgart, Metzler

Grondelaers S 2000 *De distributie van niet-anaforisch er buiten de eerste zinsplaats. Sociolexicologische, functionele en psycholinguïstische aspecten van er's status als presentatief signaal.* Unpublised PhD thesis, K.U. Leuven.

Hawkins J 1994 *A performance theory of order and constituency.* Cambridge, CUP.

Hoberg U 1981 *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache.* München, Max Hueber Verlag.

Kurz D 2000 *Wortstellungsphänomene im Deutschen.* Unpublished Master thesis, Universität Saarbrücken.

Lenerz J 1977 *Zur Abfolge nominaler Satzglieder im Deutschen.* Tübingen, Niemeyer.

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

Pechmann T, Uszkoreit H, Engelkamp J, Zerbst D 1996 Wortstellung im deutschen Mittelfeld. Linguistische Theorie und psycholinguistische Evidenz. In Habel C, Kanngießer S, Rickheit G (eds), *Perspektiven der kognitive Linguistik. Modelle und Methoden.* Opladen, Westdeutscher Verlag, 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 Rosengren I (ed), *Sprache und Pragmatik: Lunder Symposium 1986.* Stockholm, Almqvist & Wiksell International, pp 139-177.

Shannon T 2000 On the order of (pro)nominal arguments in Dutch and German. In Shannon T, Snapper J (eds), *The Berkeley Conference on Dutch Linguistics 1997*. Lanham (MD), University Press of America, pp 145–196.

Speelman D 1997 *Abundantia Verborum. A computer tool for carrying out corpus-based linguistic case studies.* Unpublished PhD thesis, K.U.Leuven.

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