# A Statistical Account on Word Order Variation in German

**Daniela Kurz**
Computational Linguistics, Saarland University
Postfach 15 11 50, 66041 Saarbrücken, Germany
kurz@coli.uni-sb.de

## Abstract

In this paper we present a corpus-based study involving the linear order of subject, indirect object and direct object in German. The aim was to examine several hypotheses derived from Hawkins' (1994) performance theory. In this context it was crucial to examine whether and to which extend length influences the order of subject and objects. The analysis was based on data extracted from the annotated NEGRA corpus (Skut et al., 1998) and the untagged Frankfurter Rundschau corpus. We developed an analysis system operating on the untagged corpus that facilitates the acquisition of data and subsequent statistical analysis. In the following, we describe this system and discuss the results drawn from the analysis of the data. These results do not support the theoretical assumptions made by Hawkins. Furthermore, they suggest the investigation of other factors than length.

## 1 Background and Motivation

Based on the assumption that basic word order regularities are reflected in the frequency of their occurrence, a corpus-based study involving word order phenomena in German was carried out. A number of different parameters have been linked with the linearization of complements and adjuncts in languages exhibiting a relatively free word order. The main factors that have been proposed are: Pronominality, case, information structure, definiteness, thematic roles, stress and length. Theories found in literature range from predominantly competence-based models to explanations almost entirely based on performance assumptions.

Recently, Hawkins' (1994) length-based theory has received much attention in general linguistics (typology), computational linguistics (modelling language evolution) and psycholinguistics (memory-based models of sentence processing). According to Hawkins, the influence of all factors other than length can almost entirely be explained as epiphenomena of length. The data presented by Hawkins are suggestive but much too restricted in size to permit any empirically supported conclusions.

In this paper we will report on a corpus-based study involving six German verbs exhibiting different basic order patterns namely, NOM<DAT, DAT<NOM, ACC<DAT, DAT<ACC . In order to reduce the effects of other factors, we restricted the investigation to non-pronominal NPs in the middle field. Two corpora were chosen for the analysis: a syntactically annotated corpus of German newspaper text, the NEGRA corpus, and the untagged Frankfurter Rundschau corpus. For each of the two verb groups (transitive and ditransitive), pairs of NOM<DAT, DAT<NOM and ACC<DAT, DAT<ACC were considered. The aim was to search for any interdependence of word order and length. Apart from this, we investigated the impact of definiteness on word order. The data acquisition and analysis were to a large part automated.

Section 2 summarizes the main ideas of Hawkins' length-based theory. In section 3 we present the statistical investigation and discuss the results in section 4. Finally we conclude with section 5.

## 2 Hawkins' Performance Theory

The theory is based on the assumption that limitations on working memory influence the construction of constituents and that humans prefer to arrange constituents in orders that minimize processing effort. Sentence processing is therefore determined by the principle of

*Early Immediate Constituents (EIC).* Hawkins assumes that phrases are constructed deterministically in a bottom-up fashion. To construct a phrasal node, it is mostly sufficient to recognize a prefix of the new phrase, so that there is no need to wait until all its immediate constituents (ICs) are found. Hawkins postulates that the prefix-based construction of new phrases is triggered by some lexical or phrasal category that uniquely identifies the mother node to be constructed. For example, German and English NPs are recognized at their left periphery, which can be a determiner, an adjective or the head noun. For German (as for English) the main claim of this theory is that all types of phrases that tend to precede their siblings, such as topic phrases, pronominal NPs, complements preceding in basic order and definite NPs are shorter on average than their respective counterparts. This system predicts that example (1) in Figure 2 will be easier to comprehend than example (2), since in the former 4 words have to be parsed instead of 11 in the latter to arrive at the constituent structure of the VP, the NP, and the PP.

In the context of the present study EIC predicts that a short nominative precedes a long dative and that a short dative precedes a long nominative. The same holds for the sequence of accusatives and datives. In addition, Hawkins assumes that verbal position interacts with length in determining the basic word order. In verb-first and verb-second sentences "short before long" should be strongly preferred, in verb-final sentences "short before long" should be slightly preferred. Hence, the following parameters were investigated in this study:

- length of the NP;
- verbal position;
- definiteness.

Due to lack of space, we will focus in what follows on parameter length and its interaction with definiteness. Kurz (2000) presents the investigations of all parameters in detail.

## 3 Statistical Investigation

Statistical analysis was performed in two steps:

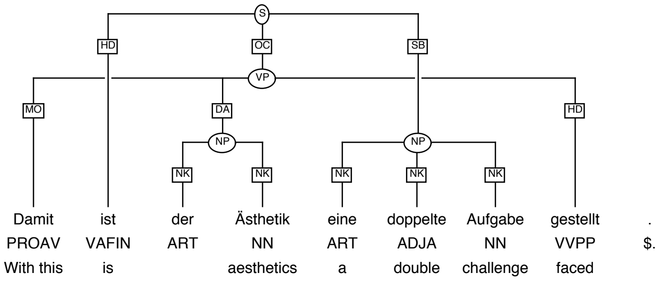
- Determining the verbs to be investigated by extracting all sentences that exhibit the relevant word pattern (NOM<DAT, DAT<NOM, ACC<DAT, DAT<ACC) from the NEGRA corpus.
- Determining the frequencies of the word patterns each verb occurs with using the much larger Frankfurter Rundschau corpus. The distribution found in this corpus was then statistically analysed.

### 3.1 Extraction of Word Patterns

In order to determine a set of interesting verbs by pursuing an empirical approach, the NEGRA corpus was chosen. The NEGRA corpus is a treebank currently consisting of 20 000 sentences or 355 000 tokens. The annotation scheme of the treebank combines phrase-structures and dependency-based schemes (cf. (Skut et al., 1997). Three types of information are encoded:

- predicate argument structure: trees with possibly crossing branches;
- syntactic categories: node labels and part-of-speech tags;
- functional categories: edge labels.

The representation format is shown in the figure[1] below:



*With this aesthetics is faced with a double challenge*

Figure 1: Encoding of a sample structure

All relevant word order patterns have been extracted by implementing matching routines operating on the structure of the corpus. The determination of the middle field was achieved by a decision tree dealing with the possible patterns displayed in Table 1.

[1] Edge labels: HD head, OC clausal object, SB subject, MO modifier, DA dative, NK noun kernel. Crossing edges indicate discontinuous constituency.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (1) | I | VP[gave | PP[to | Mary] | NP[the | valuable | book | that | was | extremely | difficult | to find]] |
|  |  | 1 | 2 | 3 | 4 |  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (2) | I | VP[gave | NP[the | valuable | book | that | was | extremely | difficult | to | find] | PP [to | Mary ]] |
|  |  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |  |

Figure 2: Recognition of phrasal categories

| **left sentence bracket** | **right sentence bracket** |
|---|---|
| finite verb | empty or separable verbal prefix |
| auxiliary or modal | verbal complex e.g. perfect participle |
| complementizer | finite verb or verbal complex |

Table 1: Possibilities of filling the sentence brackets

The positions of the left and the right sentence bracket restrict the scope in which possible NP sequences can occur. In Figure 1 the left sentence bracket consists of the finite auxiliary *ist* and the right sentence bracket is represented by the perfect participle *gestellt*. The middle field thus consists of the dative NP *der Ästhetik* and the nominative NP *eine doppelte Aufgabe*. Verbs for further investigation have been chosen according to their frequency of occurrence in each word order variation. The verbs which occurred most frequently have been selected. Tables 2 and 3 show the selected verbs and their distribution in the NEGRA corpus.

| **Verb** | **Total** | **DAT<NOM** | **NOM<DAT** |
|---|---|---|---|
| gelingen (to succeed) | 43 | 4 | 0 |
| helfen (to help) | 70 | 0 | 5 |
| zur Verfügung stehen (to be available) | 31 | 3 | 2 |

Table 2: Distribution of transitive verbs in the NEGRA corpus

| **Verb** | **Total** | **DAT<ACC** | **ACC<DAT** |
|---|---|---|---|
| geben (to give) | 560 | 16 | 0 |
| vorstellen (to present) | 38 | 0 | 3 |
| zur Verfügung stellen (to make available) | 24 | 0 | 3 |

Table 3: Distribution of ditransitive verbs in the NEGRA corpus

The second column (headed by Total) shows the numbers of all sentences containing the respective verb. In columns headed by DAT<NOM, NOM<DAT, ACC<DAT and DAT<ACC the numbers of sentences exhibiting each of the relevant patterns are given. It is evident that only a small part of the items found meets the search conditions. This is because the search was restricted to full NPs and for the most part one of the relevant NPs was pronominal.

Because of the low frequency of any individual verb in the NEGRA corpus we analysed the much larger untagged Frankfurter Rundschau corpus for the verbs under investigation.

### 3.2 Mining the Untagged Frankfurter Rundschau Corpus

The untagged Frankfurter Rundschau corpus consists of raw ASCII data. It contains 1.644 million sentences or 40.9 million tokens. Considering the size of the data it was crucial to automate the analysis as far as possible. This was achieved by developing an evaluation system making use of existing NLP tools and redefined interfaces. With this system, it was possible to ascertain, handle, and analyse the data with a minimum of manual revision.

The executed steps are shown in Figure 3.

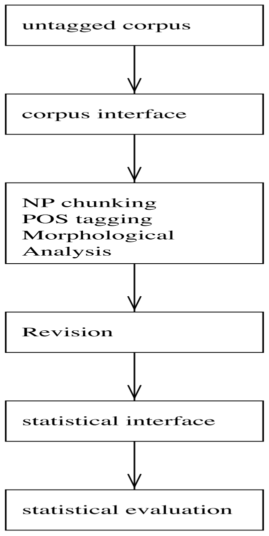


Figure 3: Flow chart of executed steps

### 3.2.1 Corpus Interface

The extraction of all sentences containing one of the verbs under investigation was done by running pattern-matching routines. These routines used regular expressions including all verbal inflections and carried out format conversions required by the subsequent component.

### 3.2.2 Shallow Parsing and Morphological Analysis

As already mentioned we were looking for particular sequences of case-marked NPs. Therefore each NP had to be labeled with case information. In order to do this, the NP boundaries had to be determined. We employed a stochastic parser (Skut's (1999) chunk tagger) that recognizes the internal structure of phrases and determines NP boundaries. As output for complex NPs, the chunk tagger delivered information about phrase boundaries and part-of-speech tags. For the annotation of structural and part-of-speech information, the chunk-tagger uses two instances of the TnT-Tagger developed by Brants (1996). The next step was followed by a morphological analysis (MORPHIX (Finkler and Neumann, 1986)) labeling each word of each NP with information about inflection based on the already available part-of-speech information. The case determination of complex NPs was done by a unification of the number, gender and case values of each word belonging to the whole NP. The output consisted of NPs labeled with partly ambiguous case information as shown in examples (3)-(5).

| | | | |
|---|---|---|---|
| (3) | [NOM ACC | Die | Gemeinde] |
| | | The | community |

| | | |
|---|---|---|
| (4) | [NOM GEN DAT ACC | Kinder] |
| | | children |

| | | | |
|---|---|---|---|
| (5) | [DAT | einem | Nachrichtenmagazin] |
| | | a | news magazine |

### 3.2.3 Revision

In the following revision, all sentences containing at least one of the relevant NPs in the middle field were automatically extracted. The case disambiguation, the determination of length and definiteness of each NP, and the determination of the verbal position was done manually. An additional program interface recoded the results of this revision in numeric variables required by the statistical interpretation.

### 3.2.4 Evaluation

For the evaluation of the system we compared its output with a manually evaluated sample containing 100 sentences and determined precision and recall. Table 4 shows precision and recall for the chunking application and the morphological analysis.

| | Precision | Recall |
|---|---|---|
| Chunk tagger | 92% | 84% |
| MORPHIX | 53% | 100% |

Table 4: Recall and precision for used NLP tools

The recall of 100% for MORPHIX has several reasons. Firstly MORPHIX is operating only on the correct output of the preceding chunk tagger; secondly, in each case the unification fails MORPHIX labels the NP with [NOM, DAT, ACC, GEN]. This perfect recall is responsible for the low precision of 53% because of the frequently occurring fourfold case assignment. However, this behaviour of MORPHIX exactly met our requirements, since we were primarily concerned with extensive data acquisition (the automatic analysis was followed by a manual evaluation anyway.)

## 4 Results

The analysis of the data was done by using cross tables, chi square tests and the analysis of means. In spite of the corpus size, the data do not permit chi square tests in some cases. In these cases, we pursued a purely descriptive approach. For the sake of space, we will describe the results of the cross tables only. We will first have a look at the distribution of NP sequences in the Frankfurter Rundschau corpus. In addition we will consider the factors length and definiteness in isolation. We will then look at the interaction of the factors holding definiteness constant.

### 4.1 Distribution in the Frankfurter Rundschau Corpus

Table 7 shows the distribution of the examined sequences for each transitive verb, Table 8 shows the frequencies of the object sequences for ditransitive verbs.

By looking at the distribution of the NP orderings there are concentrations on either one

| Verb | Basic Order | EIC-unmarked | EIC-marked | Rearrangement | EIC-unmarked | EIC-marked | Total EIC-unmarked | Total EIC-marked |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| gelingen | 257 | 58% | 21% | 3 | 100% | 0% | 58% | 20% |
| helfen | 148 | 45% | 30% | 13 | 39% | 23% | 44% | 29% |
| zur Verfügung stehen | 119 | 63% | 18% | 37 | 41% | 24% | 54% | 24% |

Table 5: Transitive verbs and EIC

| Verb | Basis Order | EIC-unmarked | EIC-marked | Rearrangement | EIC-unmarked | EIC-marked | Total EIC-unmarked | Total EIC-marked |
|---:|---:|---:|---:|---:|---:|---:|---:|---:|
| geben | 457 | 40% | 37% | 8 | 38% | 38% | 40% | 37% |
| vorstellen | 29 | 69% | 14% | 24 | 8% | 59% | 42% | 34% |
| zur Verfügung stellen | 182 | 47% | 29% | 78 | 27% | 42% | 41% | 33% |

Table 6: Ditransitive verbs and EIC

| Verb | verb frequency | DAT<NOM | NOM<DAT |
|---|---:|---:|---:|
| gelingen | 3980 | **257** | 3 |
| helfen | 5700 | 13 | **148** |
| zur Verfügung stehen | 1974 | **119** | 37 |

Table 7: Transitive verbs

| Verb | verb frequency | DAT<ACC | ACC<DAT |
|---|---:|---:|---:|
| geben | 11354 | **457** | 8 |
| vorstellen | 3694 | 29 | 24 |
| zur Verfügung stellen | 2094 | **182** | 78 |

Table 8: Ditransitive verbs

of the two possible sequences (marked boldface) for each verb. *Gelingen* and *zur Verfügung stehen* clearly favour DAT<NOM ordering while *helfen* shows the opposite preference. A similar picture is found for the ditransitive verbs, with one exception: *vorstellen*, both sequences are represented nearly the same. *Geben* and *zur Verfügung stellen* show a clear preference for DAT<ACC ordering. Given this distribution, it seems more appropriate to determine basic word order dependent on the particular verb, rather than specifying a *general* basic order of arguments (cf. (Haider, 1993)). Thus, we consider orderings exhibiting high frequencies as basic orders and orderings exhibiting low frequencies as rearrangements, e.g. the basic order of *gelingen* is DAT<NOM, the rearrangement is NOM<DAT, *helfen* appears with NOM<DAT as basic order and DAT<NOM as rearrangement. For *vorstellen*, the issue which ordering is the basic order and which ordering is the rearrangement cannot be determined from the empirical distribution, since this verb is roughly equi-based with respect to NP-order.

Against the background of Hawkins' model the question arises if both basic order and rearrangement can be motivated by length phenomena.

### 4.2 How Good Are EIC's Predictions?

Tables 5 and 6 show the total percentages of the EIC-unmarked and EIC-marked cases, the basic order and the rearrangement of each verb. The EIC-unmarked cases indicate those for which EIC makes the right predictions (short NP precedes long NP), in the EIC-marked cases EIC makes the wrong predictions (long NP precedes short NP). The proportions of the marked and unmarked cases do not add up to 100% because the cases with two NPs of equal length have not been considered.

In total we can observe that the EIC predictions are fairly good for the unmarked cases but still there is a considerable amount of cases deviating from EIC. Apart from this the EIC predictions for rearrangements are worse than for the basic order. This becomes clear from the pattern of *zur Verfügung stehen* of Table 5 and *zur Verfügung stellen* of Table 6. Comparing the EIC-unmarked and the EIC-marked cases of the basic orders with the EIC-unmarked and EIC-marked cases of the rearrangements, the proportions of the unmarked cases are lower and the proportions of the marked cases are higher in the rearrangements. The same holds for *vorstellen*. We may not determine basic order and rearrangement but the EIC predictions are quite bad for one of the two orderings (shown in the column headed by rearrangement).

Since the rearrangements of *helfen, gelingen* and *geben* are underrepresented (*gelingen*: 3, *helfen*: 13 and *geben*: 8), there is no evidence in favour or against the described observation.

| Verb | Basic Order |  |  |  | Rearrangement |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | def-indef | indef-def | def-def | indef-ndef | def-indef | indef-def | def-def | indef-ndef |
| gelingen | 36% | 2% | 62% | - | - | - | 100% | - |
| helfen | 16% | 19% | 62% | 3% | 39% | - | 55% | 8% |
| zur Verfügung stehen | 72% | - | 16% | 12% | 16% | 5% | 79% | - |

Table 9: Transitive verbs and definiteness

| Verb | Basic Order |  |  |  | Rearrangement |  |  |  |
|---|---|---|---|---|---|---|---|---|
|  | def-indef | indef-def | def-def | indef-indef | def-indef | indef-def | def-def | indef-indef |
| geben | 56% | 7% | 26% | 11% | 38% | 12% | 50% | - |
| vorstellen | 17% | 3% | 73% | 7% | 8% | - | 92% | - |
| zur Verfügung stellen | 66% | 1% | 18% | 15% | 7% | 19% | 69% | 5% |

Table 10: Ditransitive verbs and definiteness

To conclude, length phenomena do not seem to be the only reason for deviation from the basic order. The results suggest to concentrate on additional factors.

## 4.3 Definiteness

Tables 9 and 10 show the distribution of the sequences of definite and indefinite NPs in our data set.

For the basic order of both verb groups it can be observed that def-indef (definite NP precedes indefinite NP) and def-def sequences (definite NP precedes definite NP ) occur most frequently. For the rearrangements, it is striking that most of the sequences belong to the def-def pattern. This holds for the transitive as well as for the ditransitive verbs. One exception can be found: the indef-def sequences of *zur Verfügung stellen*. Since Hawkins claims that all factors apart from length are epiphenomenal, the effect of length should be even stronger if the other factors (e.g. definiteness) are held constant. In our analysis, this means that we should see a clear effect of EIC for the rearranged groups because of the high amount of def-def sequences. This, on the other hand, is at odds with the results we have already drawn from Tables 5 and 6. Recall that, EIC made worse predictions for the rearrangements than for the basic order.

## 4.4 EIC Revisited

To test the expectations mentioned above we evaluated the EIC predictions for rearranged def-def sequences. The results are listed in Tables 11 and 12.

| Verb | Rearrangement | def-def | EIC unmarked | EIC marked | Total EIC unmarked | Total EIC marked |
|---|---|---|---|---|---|---|
| gelingen | NOM<DAT | 3 | 100% | - | 58% | 20% |
| helfen | DAT<NOM | 7 | 57% | 14% | 44% | 29% |
| zur Verfügung stehen | NOM<DAT | 29 | 21% | 43% | 54% | 24% |

Table 11: Transitive verbs: EIC and definiteness

| Verb | Rearrangement | def-def | EIC unmarked | EIC marked | Total EIC unmarked | Total EIC marked |
|---|---|---|---|---|---|---|
| geben | ACC<DAT | 4 | 50% | 25% | 40% | 37% |
| vorstellen | ACC<DAT | 22 | 9% | 55% | 42% | 34% |
|  | DAT<ACC | 21 | 71% | 10% | 42% | 34% |
| zur Verfügung stellen | ACC<DAT | 54 | 30% | 41% | 41% | 33% |

Table 12: Ditransitive verbs: EIC and definiteness

The contribution of EIC does not meet the expectation derived from the "epiphenomenon hypothesis". Furthermore, the data demonstrate that EIC makes incorrect predictions for verbs exhibiting a large proportion of rearranged def-def sequences (*zur Verfügung stehen, vorstellen, zur Verfügung stellen*). For the verbs with few rearranged def-def orderings (*gelingen, helfen, geben*), EIC makes fairly good predictions. Comparing the definite, rearranged EIC-marked cases with the total percentages of the EIC marked cases (serving as baseline) the proportions of the definite, rearranged EIC marked cases are above baseline for the verbs exhibiting high frequencies. Comparing the definite, rearranged EIC-unmarked cases with the total percentages of the EIC-unmarked cases the proportions of the definite, rearranged EIC-unmarked cases are below baseline. Again, this supports the assumption that EIC does not dominate rearrangements. The results indicate that a closer examination of information-based parameters such as *Topic* and *Focus*, which are correlated with definiteness, will be required.

## 5 Conclusions

The results presented in this paper suggest that:

- The determination of basic order and rearrangement depends on the particular verb. The data do not support specification of a *general* basic order.
- EIC is not the primary factor determining the linearization of complements in the middle field.
- Considering rearranged sequences the factor definiteness dominates EIC.

The made observations are at odds with Hawkins' (1994) claiming that length is the only factor determining word order. For example definiteness determines word order even in those cases where EIC cannot motivate the ordering. Moreover, several parameters seem to interact and determine the sequence to a different extent, a claim that has already been proposed by Uszkoreit (1987).

The present results emphasize the necessity of further empirical research based on interpreted and uninterpreted corpora. Especially an examination of the interaction between definiteness and information-based factors requires further extensive corpus-based studies.

The insights gained from these methodologies show that linguists cannot rely exclusively on introspective judgements as their sole source of data. Furthermore, we hope to have demonstrated the productivity of employing corpus based studies in syntactic research.

## 6 Acknowledgements

This research was funded by the DFG Sonderforschungsbereich 378 Resource-Adaptive Cognitive Processes, Project C3 Concurrent Grammar Processing. I am grateful to Markus Becker, Stephan Oepen, Oliver Plaehn, Christoph Scheepers, Hans Uszkoreit and Feiyu Xu for constructive comments and their contributions.

## References

Thorsten Brants. 1996. TnT – A Statistical Part-of-speech Tagger. Technical report, Saarland University.

W. Finkler and G. Neumann. 1986. Morphix – ein hochportabler Lemmatisierungsmodul für das Deutsche. Technical report, Saarland University, FB Informatik.

Hubert Haider. 1993. *Deutsche Syntax - generativ.* Gunter Narr Verlag, Tübingen.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency.* CUP, Cambridge.

Daniela Kurz. 2000. Wortstellungspräferenzen im Deutschen. Master's thesis, Saarland University.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing ANLP-97*, pages 88–95, Washington, DC.

Wojciech Skut, Thorsten Brants, Brigitte Krenn, and Hans Uszkoreit. 1998. A Linguistically Interpreted Corpus of German Newspaper Text. In *Proceedings of the ESSLLI Workshop on Recent Advances in Corpus Annotation*, pages 18–24, Saarbrücken, Germany.

Wojciech Skut. 1999. *Partial Parsing for Corpus Annotation and Text Processing.* Ph.D. thesis, Saarland University.

Hans Uszkoreit. 1987. *Word Order and Constituent Structure in German*, volume 8 of *Lecture Notes.* CSLI, Stanford.