How flexible is constituent order in the midfield of German subordinate clauses? A corpus study revealing unexpected rigidity

Gerard Kempen[1] and Karin Harbusch[2]
[1] Psychology Dept., Leiden Univ. & MPI for Psycholinguistics, Nijmegen
[2] Computer Science Dept., Univ. of Koblenz–Landau
kempen@fsw.leidenuniv.nl, harbusch@informatik.uni-koblenz.de

It is almost a commonplace that word order in the midfield of German clauses is flexible. Although statements to this effect do not claim that "anything goes", they suggest that word order variability in German clauses is considerably greater than, for example, in Dutch and English clauses. Few systematic empirical studies of the actual amount of variation that go beyond the intuition of the individual linguist have been published as yet. In the present paper, we adduce empirical data drawn from a corpus study on the linear order of Subject (SB), Indirect Object (IO) and Direct Object (DO) in German subordinate clauses.

In recent psycholinguistic work, word order intuitions have been probed in a systematic fashion. Keller (2000) elicited "gradient grammaticality" judgments via a novel technique based on the psychophysical method of Magnitude Estimation (Bard *et al.*, 1996). One of his goals was to assess the relative strength of three well–known NP ordering constraints (cf. Uszkoreit, 1987; Pechmann *et al.*, 1996; Müller, 1999):
C1: Pronominal $\prec$ Non–pronominal
C2: Nominative $\prec$ Non–nominative, and
C3: Dative $\prec$ Accusative
(where "$\prec$" means "precedes"). Keller found that C1 and C2 were about equally strong and both stronger than C3. This in contrast with a prediction from Müller (1999) who classified constraint C1 as more powerful than C2 and C3. Most importantly, none of the three constraints was "absolute" in the sense that its violation gave rise to an extremely low acceptability/grammaticality judgment (as low as violation of the absolute verb-final constraint in subordinate clauses did).

Our aim was to investigate to what extent informally or formally obtained linguistic grammaticality judgments are mirrored by the frequency of the various orderings in sentence materials generated outside the laboratory.

Recently, the NEGRA–II corpus has become available — a German treebank containing about 20,000 newspaper sentences annotated in full syntactic detail (Skut *et al.*, 1997). Using version 2.1 of TIGERSearch (König and Lezius, 2000), we extracted all finite clauses introduced by a subordinating conjunction and containing an (SB,IO) and/or and (SB,DO) pair, possibly with an additional (IO,DO) pair (with the members of a pair occurring in any order). As for terminology, clauses containing only an (SB,IO) pair are labeled *in*transitive; clauses with only an (SB,DO) pair are *mono*transitive; a clause with an (SB,IO) as well as an (IO,DO) pair is *di*transitive; both latter types of clauses are called transitive. We found 907 monotransitive, 99 intransitive, and 54 ditransitive subclauses meeting the requirements. We distinguished six types of pronominal and full (non–pronominal) NPs: SBpro, SBful, IOpro, IOful, DOpro and DOful. An NP is pronominal iff it consists of a personal or a reflexive pronoun. As a clause contains at most one token of each of the three types of grammatical function, there are 12 possible *unordered pairs* of NPs: three combinations of grammatical functions ((SB,IO), (SB,DO) and (IO,DO)) times four combinations of NP shapes (all pronominal, one member full, the other member full, all full). For each of these, we determined the frequency of the two possible orderings (i.e., of 24 *ordered pairs*).

There were 1168 ordered pairs: one from each mono– or intransitive clause; three from every ditransitive clause (Table 1). The first thing to be noted is the high proportion of (almost) empty cells. This suggests that grammaticality judgments tend to be more lenient than frequencies: Quite a few orderings that are rated at least average in quality, do not occur in actual practice. Stated differently, the actually observable orderings cluster in the upper regions of the grammaticality spectrum. Consequently, the level of flexibility emerging from the frequency counts is considerably lower than grammaticality intuitions suggest. The ordering of the pronominal constituents is invariably SBpro—DOpro—IOpro, and SB is the only full NP that may precede a pronominal NP; and SBful never interrupts a sequence of pronominal constituents. Variability within full NPs is somewhat greater: While SBful—IOful—DOful is the predominant order, inversions do occur regularly. Closer inspection of the data reveals, however, that the inverted order IOful—SBful is restricted to clauses with intransitive verbs (more precisely, "experiencer–object" verbs as in *daß [dem Jungen]$_{IO}$ etwas$_{SB}$ widerfährt*, 'that [the boy] something happens–to'; that something happens to the boy), and that the sequence DOful—IOful only occurs as standard order licensed by special ditransitive verbs (cf. *jemanden$_{DO}$ [einer Prüfung]$_{IO}$ unterziehen*, 'someone [a test] subject–to'; subject someone to a test). Furthermore, grammaticality ratings and frequencies are correlated insofar as the best rated ordering of the members of a constituent pair also tends to be the most frequent one.

The frequency data can be accounted for by the rather rigid rule schema in Figure 1. To each individual constituent, the schema assigns a standard ("primary") position before or after its clausemates. Each of the full NPs has a single "secondary" placement option, which is indicated by the labeled arrows. This is "freedom in restraint", con-

Table 1: Frequency of the 24 ordered pairs of grammatical functions in clauses extracted from the NEGRA–II corpus. Dark gray cells represent impossible constituent pairs.

		Second NP						*Total*
		SBpro	DOpro	IOpro	SBful	IOful	DOful	
First NP	SBpro	■	**53**	**17**	■	**21**	**246**	*337*
	DOpro	**0**	■	**1**	**120**	**10**	■	*131*
	IOpro	**0**	**0**	■	**29**	■	**31**	*60*
	SBful	■	**63**	**7**	■	**59**	**478**	*607*
	IOful	**0**	**0**	■	**20**	■	**9**	*29*
	DOful	**0**	■	**0**	**1**	**3**	■	*4*
Total		*0*	*116*	*25*	*170*	*93*	*764*	*1168*

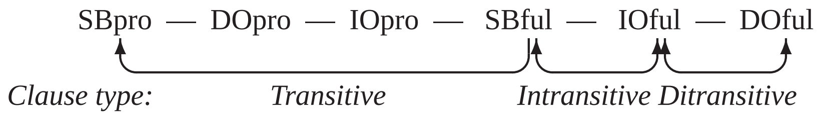

Figure 1: Rule schema representing the linearization options observed in the treebank in clauses headed by a mono–, di–, or intransitive head verb.

ditional upon mono–, di–, or intransitivity of the head verb. Mild conceptual factors such as animacy (Kempen and Harbusch, 2004a), definiteness (Kurz, 2000) and referential ease (yielding short and simple NPs; Hawkins, 1994; Wasow, 2002; Kempen and Harbusch, 2004b) enable full constituents to occupy the secondary, more leftward position. Since the schema only generates orderings that tend to elicit (relatively) high grammaticality ratings, it seems reasonable to view it as representing all and only the *unmarked* orderings. Presumably, the constraints imposed by the schema can only be overruled by strong conceptual influences related to, e.g., topic/focus relationships or complex reference, thereby giving rise to marked ("tertiary") ordering patterns not covered by the schema. However, this will be a relatively rare occasion. (One should keep in mind that the conclusions drawn so far have emerged from a relatively small corpus and need cross–checking against a larger and more varied collection of texts.)

If a rule schema such as the one in Figure 1 indeed underlies ordering decisions in subordinate clauses, this would explain why none of the three above constraints deserves the epitheton "absolute" or "hard" (Keller, 2000): The schema includes systematic exceptions of each of them. In case of a pronominal DO, for example, SBful violates Constraint C1 (Pronominal $\prec$ Non–pronominal) if it selects its secondary option; if it opts for its primary position, C2 (Nominative $\prec$ Non–nominative) is violated; and C3 (Dative $\prec$ Accusative) does not apply at all, irrespective whether IO is pronominal or full.

The schema in Figure 1 is more successful because it considers grammatical function *and* syntactic shape simultaneously. An important additional advantage of the rule schema is that it enables grammatical constituents to select their primary or secondary slot on a first-come first-serve basis, without the need to take properties of clausemates into account. This is because the schema allots absolute rather than relative positions.

Certain recent approaches to linearization make use of so–called *topological fields* (see Kathol (2000) in particular). In our work on the Performance Grammar formalism (Harbusch and Kempen, 2002; Kempen and Harbusch, 2002), we have proposed "topologies" for German clauses that consist of nine ordered slots:

F1	M1	M2	M3	M4	M5	M6	E1	E2

The slot labeled F1 makes up the forefield (from German *Vorfeld*); the slots M1 through M6 belong to the midfield (*Mittelfeld*); E1 and E2 define the endfield (*Nachfeld*). To–be–ordered constituents select a slot depending on their grammatical function and syntactic properties. For instance, slot M1 serves as the "landing site" for the finite verb of main clauses. In subordinate clauses, M1 is occupied by the subordinating conjunction (if present), while the finite verb goes to M6; and M2 and M3 each can accommodate one or more non-Wh argument NPs.

How can we fit the six–slot rule schema of Figure 1 into the nine–slot clausal topology? Somewhat arbitrarily, and in order to avoid inconsistency with some earlier publications, we propose to divide slots M2 and M3 into three subslots. This yields six receptacles — one for each of the six NP types of the rule schema (see Figure 2).

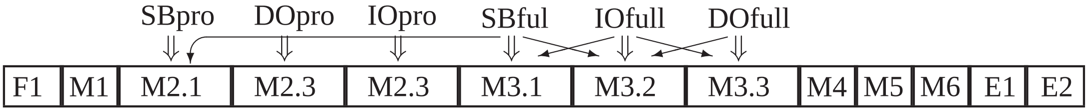

Figure 2: Mapping of argument NP types onto midfield (sub)slots. Double arrows: primary placement options; single arrows: secondary options.

The corpus frequencies we observed seem to originate from more complex and less flexible linearization constraints than is standardly assumed. The constraints do not involve *single* features such as exemplified by C1 through C3 above but *combinations* of features, in particular grammatical function and syntactic shape. Furthermore, the constraints suggest a linearization system where individual constituents receive *absolute* rather than *relative* positions. The corpus data may thus provide indirect support for recent "topological" approaches to linear order in linguistics and psycholinguistics.

References

Bard, E., D. G., Robertson, and A. Sorace (1996). Magnitude estimation of linguistic acceptability. *Language*, **72**:32–68.

Harbusch, K. and G. Kempen (2002). A quantitative model of word order and movement in English, Dutch and German complement constructions. In *Procs. of the 19th COLING-2002*. Taipei, Taiwan.

Hawkins, J. A. (1994). *A Performance Theory of Order and Constituency*. Cambridge University Press, Cambridge, U.K.

Kathol, A. (2000). *Linear Syntax*. Oxford University Press, New York, NY.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. Unpublished Ph.D. thesis, Univ. of Edinburgh, U.K.

Kempen, G. and K. Harbusch (2002). Performance Grammar: A declarative definition. In A. Nijholt, M. Theune, and H. Hondorp, eds., *Computational Linguistics in the Netherlands 2001*. Rodopi, Amsterdam, The Netherlands and New York, NY.

Kempen, G. and K. Harbusch (2004a). A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In T. Pechmann and C. Habel, eds., *Multidisciplinary approaches to language production*. Mouton De Gruyter, Berlin.

Kempen, G. and K. Harbusch (2004b). Generating natural word orders in a semi–free word order language: Treebank–based linearization preferences for argument NPs in subordinate clauses of German. In A. Gelbukh, ed., *Procs. of the Fifth CICLING, Seoul, Korea*; Lecture Notes in Computer Science, Springer, Berlin.

König, E. and W. Lezius (2000). A description language for syntactically annotated corpora. In *Procs. of the 18th COLING*, Saarbrücken.

Kurz, D. (2000). A statistical account on word order variation in German. In A. Abeillé, T. Brants, and H. Uszkoreit, eds., *Procs. of the COLING Workshop on Linguistically Interpreted Corpora*, Luxembourg.

Müller, G. (1999). Optimality, markedness, and word order in German. *Linguistics*, **37**:777–815.

Pechmann, T., H. Uszkoreit, J. Engelkamp, and D. Zerbst (1996). Wortstellung im deutschen Mittelfeld. Linguistische Theorie und psycholinguistische Evidenz. In *Perspektiven der Kognitiven Linguistik*. Westdeutscher Verlag, Wiesbaden.

Skut, W., B. Krenn, T. Brants, and H. Uszkoreit (1997). An annotation scheme for free word order languages. In *Procs. of the Fifth ANLP*, Washington D.C.

Uszkoreit, H. (1987). *Word Order and Constituent Structure in German*. CSLI Publication, Stanford, CA.

Wasow, T. (2002). *Postverbal behavior*. CSLI Publications, Stanford, CA.