Chapter
A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment

Gerard Kempen and Karin Harbusch

1. Introduction

The grammar of German does not impose hard constraints on the linear order of Subject (SB), Indirect Object (IO) and Direct Object (DO) in finite complement or adverbial clauses (for an overview of the linguistic literature see Müller (1999)). All six possible orders are acceptable, although with varying degrees of grammaticality (Keller 2000). Given this flexibility, which factors control the actual linearization preferences of speakers/writers of German?

Many studies into this question have proposed linear precedence rules in syntactic terms, e.g. SB ≺ IO/DO; pronominal NPs ≺ full NPs; IO ≺ DO (with the symbol "≺" means "precedes"; cf. Uszkoreit 1987; Pechmann, Uszkoreit, Engelkamp, and Zerbst 1996; Müller 1999). Other studies have explored the impact of conceptual factors, e.g. whether the NP fulfilling a given grammatical role is definite or indefinite (Kurz 2000), and whether it refers to an animate or inanimate entity (Dietrich and van Nice, in press).

The literature offers two suggestions as to how animacy could exert its influence on the linearization process (McDonald, Bock, and Kelley 1993; Feleki and Branigan 1997; Branigan and Feleki 1999; Dietrich and van Nice, in press). The first possibility is an *indirect* influence: Animacy could affect the assignment of grammatical functions to an NP. Many verbs require their Subject to play the thematic role of "agent". Since animate NPs are prototypical agents, they could have a higher probability of becoming Subject of the clause than inanimate NPs. Precedence rules such as "SB ≺ IO/DO" enable Subject NPs to occupy an early position in the clause. The second suggestion, which presupposes incremental sentence production, is a *direct* influence of animacy on constituent order. NPs with animate referents tend to be conceptualized, assigned thematic and functional roles, and attached to the surface structure prior to inanimate NPs. In a language with relatively free word order, this may occasion animate NPs to receive a position leftward of an inanimate NP, independently of their grammatical function (Branigan and Feleki 1999; Kempen and Harbusch 2003). For instance, if the grammar of the target language leaves the order of SB and IO unspecified, animacy could be one of the factors determining whether SB or IO will lead. It goes without saying that the two theoretical suggestions do not rule out one another.

The psycholinguistic literature contains several experimental sentence production studies evaluating the "direct" and the "indirect" hypotheses. McDonald, Bock, and Kelly (1993) observed, with English speaking participants, that animacy significantly affects the assignment of grammatical functions — in line with the "indirect" hypothesis. Branigan and Feleki (1999) report production data, obtained with Greek participants, in support of the "direct" hypothesis. Experimental studies with German participants yielded support for both hypotheses (see Dietrich and Van Nice, in press, and the references therein).

Recently, the NEGRA-II corpus (Skut et al. 1997) has become available, containing about 20,000 newspaper sentences annotated in full syntactic detail (see Figure 1 for an example). This prompted us to conduct the first corpus study into the issue at hand on the basis of sentence materials produced outside the laboratory.

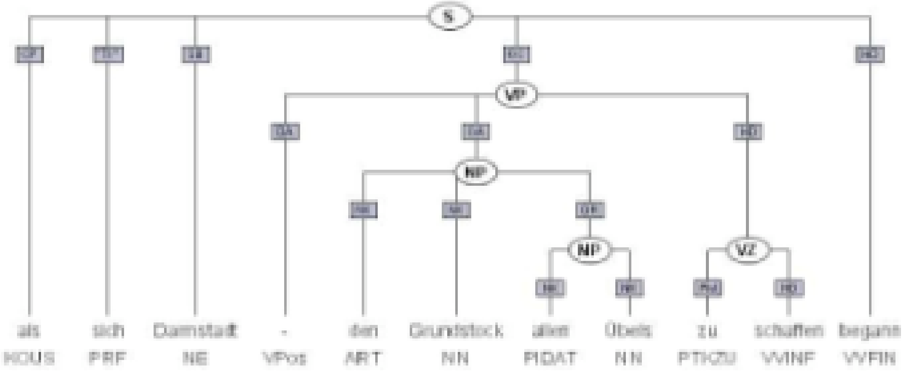

Figure 1. Example of the NEGRA-II annotation for a subordinate clause.

2. Method

From the NEGRA-II treebank, we extracted all adverbial and complement clauses containing (SB,IO) and/or (SB,DO) pairs, possibly with an additional (IO,DO) pair (see Appendix for details of the extraction method). As for terminology, clauses containing only an (SB,IO) pair are called intransitive. We distinguish two types of transitive clauses: those including only an (SB,DO) pair are termed monotransitive; clauses containing three pairs — (SB,DO), (SB,IO) as well as (IO,DO) — are ditransitive. We found 907 monotransitive, 99 intransitive, and 54 ditransitive finite subclauses. Because, as is well-known, pronominal and full (i.e. non-pronominal) NPs are linearized according to different rules, we distinguished six types of NPs: SBpro, SBful, IOpro, IOful, DOpro and DOful. An NP is pronominal if it consists of a personal or a reflexive pronoun. As a clause contains at most one token of each of the three types 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, only first member full, only second member full, all full). For each of these, we determined the frequency of the two possible orderings (i.e., of 24 *ordered pairs*; see Table 1).

Each member NP of a pair was classified as animate/inanimate, definite/indefinite and pronominal/full. Animacy was defined as "referring to a human or an animal, or a collective of humans/animals". In case of doubt ("animate?"), we counted a referent as animate. Reflexive pronouns received the same animacy value as their antecedents. (There were no reciprocal pronouns fulfilling SB, IO or DO function.) NPs headed by a personal (*ich* 'I', *er* 'he', *uns* 'us' *sie* 'they/them', etc.) or reflexive pronoun (*sich(selbst)* 'himself/herself/themselves', *dich(selbst)* 'yourself') were considered pronominal (all other pronouns, including the indefinite pronoun *man* 'one', were counted as a full NP). For reasons to be explained below, we also determined whether the NPs were definite or indefinite. As definite we considered proper names and NPs that contain a definite determiner (including a demonstrative or possessive pronoun or a possessive genitive NP) or are headed by a personal pronoun. In case of a reflexive pronoun, we assigned a definiteness value identical to that of its antecedents.

3. Results

Table 1 shows the observed orderings of all 1168 function pairs extracted from the corpus: one pair from each of the intransitive and monotransitive clauses, three pairs from every ditransitive clause. It reveals that no less than10 out of 12 pairs manifest a strong bias to occur in one of the two possible orderings; the order of their member NPs is more or less fixed. We found only two pair types where both NP orders occur with sufficiently high frequency to warrant testing the influence of animacy. These pairs are (DOpro,SBful) and (SBful,IOful); see Tables II and III for the distribution of (in)animacy among their member NPs.

In (DOpro,SBful) pairs, animacy of the SBful member significantly increases the likelihood for this member to precede the other member ($\square^2$=23.5; p<.0005; see Table 2). The position of the IOful member of (SBful,IOful) pairs is affected similarly ($\square^2$=17.8; p<.0005; see Table 3). Both tendencies cannot be attributed to definiteness of the NPs because animacy and definiteness turn out to be uncorrelated in the relevant NPs: For the (SBful,DOpro) pairs, the phi correlation coefficient $r_{\square}$=.10; $\square^2$=.22; n.s); for the (SBful,IOful) pairs $r_{\square}$=.21; $\square^2$=.02; n.s).

Table 1. Frequencies of the 12 possible pairs of grammatical functions in finite subordinate clauses extracted from the NEGRA-II corpus.

Preferred order of NP pairs	Frequency of preferred order	Frequency of opposite order
(SBpro,DOpro)	53	0
(———,IOpro)	17	0
(———,IOful)	21	0
(———,DOful)	246	0
(DOpro,IOpro)	1	0
(———,SBful)	120	63
(———,IOful)	10	0
(IOpro,SBful)	29	7
(———,DOful)	31	0
(SBful,IOful)	59	20
(———,DOful)	478	1
(IOful,DOful)	9	3

Table 2. Linear order frequencies of (SBful,DOpro) pairs, for animate and inanimate SBful members.

	Linear order	
	SBful $\prec$ DOpro	DOpro $\prec$ SBful
SBful inanimate	11	64
SBful animate	52	56

Table 3. Linear order frequencies of (SBful,IOful) pairs, for animate and inanimate IOful member*s*

	Linear order	
	IOful $\prec$ SBful	SBful $\prec$ IOful
IOful inanimate	3	39
IOful animate	17	20

4. Discussion

The corpus data strongly confirm earlier experimental results arguing for a *direct* influence of animacy on linearization. Apparently, animacy not only affects the functional structure of the clause under construction (the "indirect hypothesis") but also its linear structure. Parenthetically, the "direct hypothesis" does not entail the prediction that animate NPs always or mostly precede inanimate ones. This is because animacy is not the only factor determining the actually observable order of NPs. For example, as witnessed by Table 1, German has a strong tendency for pronominal NPs to precede full NPs. Animacy works against this force but apparently cannot cancel it out completely.

The results of our corpus study have potentially important implications in regard of the architecture of the grammatical encoding process. It is standardly assumed that the grammatical encoding process comprises two stages: Grammatical functions are assigned to constituents during the *functional* stage; their linear order is determined during a later *positional* stage (Garrett 1975, 1980; Bock and Levelt 1994; Pickering, Branigan, and McLean 2002). On this account, the indirect effect of animacy takes place during the functional stage, the direct effect during the positional stage. If so, animacy would influence grammatical encoding *twice*. This raises the question of whether a more parsimonious and therefore preferable model would be feasible where the functional and linear structures of a clause are computed in one stage. We could assume, for each constituent, that the assignment of its grammatical function and the determination of its linear position are closely interlinked, virtually simultaneous decisions (see Kempen and Harbusch (2002) for a computational model of this sort). The earlier availability of animate concepts in comparison with inanimate ones could then affect their lexicalization, functional role assignment and linear structure "in one go". Such as proposal requires rethinking the original empirical evidence for the stage models of sentence production, which however is beyond the scope of this Chapter.

Appendix: Extraction method

We searched the NEGRA II corpus using the TIGERSearch system (König and Lezius 2000), version 2.0. The query

"(#n1:[cat="S"] >CP #n2: [pos="KOUS"|cat="CCP"])"

searches for a node (addressed by variable #n1) labeled with the nonterminal feature cat (category) S. This node should yield an edge labeled CP (complementizer) linked to the node addressed by the variable #n2. Node #n2 has to be a lexical anchor of class KOUS (terminal feature pos; KOUS=subordinating conjunction) or a complex CP (nonterminal feature cat = CCP). The query retrieved 2393 subordinate clauses that had to be inspected in more detail.

Within these clauses, we searched for the grammatical functions:

- SB, i.e. subject,
- DA, i.e. dative object (in this paper referred to as indirect object, IO) and
- OA, i.e. accusative object (here referred to as direct object, DO).

As the edge label SB is always a sibling of the CP, the query becomes extended by "& (#n1 >SB #n3)", i.e., S has also to directly dominate the subject. This restriction removed 50 matches.

For the two object types DO and IO, more detailed patterns had to be defined because auxiliaries and modals yield structures nested under OC edges (Clausal Object — possibly with traces, if constituents have moved out of the scope of the dominance structure of its verb: cf. "clause union"; for an example, see the node labeled "T1" above the pronoun sich in Figure 1). In order to license all these options, the query for a DA looks as follows:

```
"// SB and DA are siblings for full verbs
& ((#n1 >DA #n4) |
// DA may be below OC (with or without traces),
```

// furthermore, it may be complex, e.g., a coordinating conjunction or a nested OC construction

((#n1 >OC #n5) & (#n5 >1,2 #n6) & (#n7 >DA #n6)))"

Notice that this query slightly overgenerates. It also retrieves clauses containing an extraposed infinitival clause without a subject (e.g., *da die Planer beabsichtigten, dem Bauherrn zusätzlich die Schaffung von Wohnungen abzuhandeln* or *als der Landrat beschloß, die asbesthaltigen Baustoffe [...] schon in den Sommerferien beseitigen zu lassen*). These sentences were removed by hand.

Another class of sentences which we did not want to consider are comparisons that embody main clause word order (e.g., *als würde mich alle vier Minuten ein Laster [...] überrollen* or *als wollte Berti Vogts dem Gruppen-Gegner vom heutigen Freitag Angst mit dem Eßlöffel einflößen*). Furthermore, a few sentences with wrong annotations were erased (e.g., *daß ihn nur mehr als die Hälfte der 36 FDP-Mitglieder in den westlichen Stadtteilen [...] zu ihrem Boß gewählt hat*, where ihn was coded as DA instead of OA).

The resulting list of sentences was converted to EXCEL format, keeping the grammatical functions SB, IO and DO as well as their relative positions. The NP features animacy/animacy?/inanimacy, definiteness/indefiniteness, and pronominal/full were handcoded. The ordering patterns of the constituents were counted by means of simple EXCEL macros (rather than via TIGERsearch).

References

Bock, J. Kathryn and Willem J.M.Levelt

1994 Language production: Grammatical encoding. In: Morton A. Gernsbacher (Ed.), *Handbook of psycholinguistics.* San Diego: Academic Press.

Branigan, Holly P., and Elina Feleki

1999 Conceptual accessibility and serial order in Greek language production. *Proceedings of the 21st Conference of the Cognitive Science Society (Vancouver).*

Dietrich, Rainer, and Kathy Y. van Nice

in press Belebtheit, Agentivität und inkrementelle Satzproduktion. In: Christopher Habel, and Thomas Pechmann (Eds.), *Sprachproduktion.* Wiesbaden: Westdeutscher Verlag.

Garrett, Merrill F.

1975 The analysis of sentence production. In: Gordon H. Bower, (Ed.), *The psychology of learning and motivation* (Vol. 9). New York: Academic Press.

Garrett, Merrill F.

1980 Levels of processing in speech production. In Brian Butterworth (Ed.), *Language production (Vol. 1).* London: Academic Press.

Feleki, Elina, and Holly P. Branigan

1997 Animacy effects in Greek sentence production: Evidence for single-stage syntactic processing? Poster presented at the Third Conference of Architectures and Mechanisms of Language Processing (AMLaP-97), Edinburgh.

Keller, Frank

2000 *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality.* Unpublished Ph.D. Thesis, University of Edinburgh.

Kempen, Gerard, and Karin Harbusch

2002 Performance Grammar: A declarative definition. In: Anton Nijholt, Mariët Theune, and Hendri Hondorp (Eds.), *Computational Linguistics in the Netherlands 2001.* Amsterdam: Rodopi.

Kempen, Gerard, and Karin Harbusch

2003 Word order scrambling as a consequence of incremental sentence production. In: Holden Härtl, and Heike Tappe (Eds.), *Mediating between concepts and grammar.* Berlin: Mouton De Gruyter.

Kurz, Daniela

2000 A statistical account on word order variation in German. In: Anne Abeillé, Thorsten Brants, and Hans Uszkoreit (Eds.), *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora, Luxembourg.*

König, Esther, and Wolfgang Lezius

2000 A description language for syntactically annotated corpora. *Proceedings of the International Conference on Computational Linguistics (COLING*), Saarbrücken, Germany.

McDonald, Janet L., J. Kathryn Bock, and Michael H. Kelly

1993 Word and world order: Semantic, phonological and metrical determinants of serial position. *Cognitive Psychology, 25,* 188-230.

Müller, Gereon

1999 Optimality, markedness, and word order in German. *Linguistics, 37,* 777-818.

Pechmann, Thomas, Hans Uszkoreit, Johannes Engelkamp, and Dieter Zerbst

1996 Wortstellung im deutschen Mittelfeld. Linguistische Theorie und psycholinguistische Evidenz. In: Christopher Habel, Siegfried Kanngießer, and Gert Rickheit (Eds.), *Perspektiven der kognitiven Linguistik.* Wiesbaden: Westdeutscher Verlag.

Pickering, Martin J., Holly P. Branigan, and Janet F. McLean

2002 Constituent structure is formulated in one stage. *Journal of Memory and Language, 46,* 586-604.

Skut, Wojciech, 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), Washington D.C*

Uszkoreit, Hans

1987 *Word order and constituent structure in German.* Stanford CA: CSLI Publications.