Generating natural word orders in a semi–free word order language: Treebank-based linearization preferences for German

Gerard Kempen[1] and Karin Harbusch[2]

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

Abstract. We outline an algorithm capable of generating varied but natural sounding sequences of argument NPs in subordinate clauses of German, a semi-free word order language. In order to attain the right level of output flexibility, the algorithm considers (1) the relevant lexical properties of the head verb (not only transitivity type but also reflexivity, thematic relations expressed by the NPs, etc.), and (2) the animacy and definiteness values of the arguments, and their length. The relevant statistical data were extracted from the NEGRA–II treebank and from hand-coded features for animacy and definiteness. The algorithm maps the relevant properties onto "primary" versus "secondary" placement options in the generator. The algorithm is restricted in that it does not take into account linear order determinants related to the sentence's information structure and its discourse context (e.g. contrastiveness). These factors may modulate the above preferences or license "tertiary" linear orders beyond the primary and secondary options considered here.

1 Introduction

Computational sentence generators should be able to order constituents in agreement with linearization preferences and habits of native speakers/writers. This knowledge can be attained by exploiting text corpora (cf. [1]). In the following we concentrate on extracting appropriate word order rules for German, a (semi-)free word order language.

Target languages with strict word order rules do not present much of a problem here although the grammaticality contrast between examples such as *Pat picked a book up* and *?Pat picked a very large mint–green hardcover book up* [2, p. 7] shows that, even in English, knowledge of linear order preferences comes in handy. In the case of (semi–)free word order languages, the problem of how to select natural sounding permutations of constituents from among those licensed by the grammar is much more widespread. Sentence generators striving for natural and varied output, e.g., in question–answering systems or computer–supported language training environments, should neither select the same permutation at all times, nor produce the various grammatical permutations at random.

The naturalness of a particular ordering of constituents often depends on subtle conceptual or pragmatic factors that grammars of the target language fail to capture. A well–known case in point is German, whose grammar does not impose hard constraints on the linear order of Subject (SB), Indirect Object (IO) and Direct Object (DO) in finite subordinate clauses. (For an overview of the relevant linguistic literature, see [3].) All six possible orders are acceptable, although with varying degrees of grammaticality [4]. Given this flexibility, which factors control the actual linearization preferences of speakers/writers of German? In this paper, we explore the feasibility of extracting relevant linear order constraints from a treebank, *in casu* the NEGRA II corpus of German [5].

Students of constituent order in German have proposed linear precedence rules such as (1) SB $\prec$ IO/DO, (2) pronominal NPs $\prec$ full NPs, and (3) IO $\prec$ DO (where the symbol "$\prec$" means "precedes"; cf. [6], [7], [3]). However, these rules are not very helpful in designing a sentence generator: As will become clear hereafter, there are systematic exceptions to each of them, and important argument ordering preferences are linked to lexical properties of the head verb. Other studies have explored the impact of conceptual factors, e.g. whether the argument NP is definite or indefinite [8], and whether it refers to an animate or inanimate entity [9]). Another factor likely to play a role is length (cf. "heavy NP shift"; [10], [2]).

In this paper we take the following determinants into consideration:

- Grammatical function: SB, IO, DO
- Form: pronominal (consisting of a personal or reflexive pronoun) or full (otherwise)
- NP length: number of terminal nodes dominated by the NP (as determined by the TIGERSearch tool [12])
- Animacy: referring to a human or animal, or a collective of humans/animals (hand–coded[3])
- Definiteness: definite vs. indefinite reference (hand–coded according to Tables 1 and 2 in [11]).

We study the influence of these factors and some their interactions in *subordinate clauses introduced by a subordinating conjunction* (for example, *daß/dass* 'that', *ob* 'whether', *weil* 'because', *obwohl* 'although', *wenn* 'when, if'). The main reason for this restriction is a strategic one: The linear ordering patterns in these subordinate clauses are simpler than those in other clause types (e.g., no obligatory fronting of Wh-constituents).

2 Method

Recently, the NEGRA–II corpus has become available, a German treebank containing about 20,000 newspaper sentences annotated in full syntactic detail. Us-

[3] In case of doubt, 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.)

ing version 2.1 of TIGERSearch, we extracted all 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). For details of the clause selection method, see [9]. 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 clauses meeting our criteria. Every argument NP in these clauses was assigned a value on each of the five properties listed above.

As noted in an earlier paper [13], the observed constituent order frequencies can be accounted for in terms of the rather rigid rule schema in Fig. 1, which assigns to individual constituents a standard ("primary") position before or after their clausemates. Full NPs have an alternative ("secondary") placement option indicated by the labeled arrows. Animacy is one of the factors determining whether or not the secondary placement option is taken [9]: In transitive clauses, full Subject NPs ("SBful") are more likely to precede pronominal DO NPs ("DOpro") if they are animate than if they are inanimate; in intransitive clauses, animate IOful NPs precede SBful significantly more often than inanimate ones do. (For the inversion of DOful and IOful, see below.)

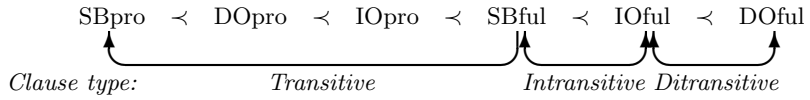

Fig. 1. Rule schema representing the linearization options observed in the treebank in clauses headed by a mono–, di–, or intransitive head verb.

3 Results

We now present new results regarding factors determining the choice between primary or secondary placement options licensed by the rule schema, as well as on some of their interactions.

Monotransitive clauses. Of 179 clauses with an (SBful,DOpro) pair, 143 have a reflexive head verb. In these clauses, *sich* 'him–, her–, itself; themselves' is the obligatory reflexive pronoun. As this pronoun is coreferential with the Subject, the members of a pair are either both animate or both inanimate. Of 65 inanimate pairs, SBful takes the secondary placement option (i.e., before DOpro) in 9 cases (14%); in the remaining 78 instances of animate pairs, SBful precedes DOpro 38 times (49%). In the remaining 36 *non–reflexive* clauses, the latter fifty–fifty pattern holds uniformly for all four possible pairings of an animate and/or an inanimate NP. Early SBs tend to be shorter than late SBs in reflexive

as well as non-reflexive clauses (mean lengths 2.62 vs. 4.03 words). In the 179 clauses considered here, definiteness of SBful has no influence on its being placed early or late: In the SBful NPs that precede the DOpro NPs, the proportion of definites is the same as in the SBful NPs that follow DOpro (65% in both cases).

Intransitive clauses. Here we distinguish three types of IOful NPs based on the thematic role they express:

- recipient: in passive clauses headed by a ditransitive verb (e.g., *übertragen werden* 'to be transferred'; IOful first: 9 clauses, SBful first: 12)
- patient (or "co–agent"): in active clauses headed by verbs such as *helfen* 'help', *folgen* 'follow', *beitreten* 'join' (IOful first: 0 clauses, SBful first: 17)
- experiencer: in active clauses headed by verbs like *gefallen* 'please', *gehören* 'belong', *entsprechen* 'correspond' (IOful first: 11 clauses, SBful first: 13).

Patient–IOful NPs (*helfen*–type) never precede the SBful NP. Experiencer–IOs (*gefallen*–type) and recipient–IOs (*übertragen werden*) all adhere to the rule "animate $\prec$ inanimate". Neither length of the NPs nor their (in)definiteness seem to play a prominent role here. Where the animacy rule does not apply, the recipient-IOs select the primary or secondary position more or less at random, whereas the experiencer–IOs invariably choose the primary position.

Ditransitive clauses. The NP orderings agree with the primary positions depicted in Fig. 1, with two exceptions. Animate SBful NPs (length ≤ 2) precede pronominal arguments with proportions roughly comparable to those in monotransitive clauses. Three clauses instantiate the inverted DOful $\prec$ IOful sequence. They contain verbs where this order is standard (e.g., *etwas*$_{DO}$ *etwas*$_{IO}$ *angleichen* 'assimilate something$_{DO}$ to something$_{IO}$').

Definiteness, animacy, and length. The observed tendency for animate arguments to precede inanimate ones cannot be attributed to definiteness of the NPs because animacy and definiteness turn out to be uncorrelated. This preference cannot be attributed to length either despite that, on average, animate NPs are shorter than inanimate ones, and short NPs also prefer early positions. The stronger leftward tendency of animate in comparison with inanimate NPs remains clearly visible if one only looks at NPs of equal length (length = 1, length = 2, or length > 2).

4 Conclusion

An algorithm capable of generating varied but natural sounding sequences of argument NPs in subordinate clauses of German can take the primary positions in the rule schema of Fig. 1 as starting point. In order to attain output flexibility, it should consider (1) the relevant lexical properties of the head verb (not only transitivity type but also reflexivity, thematic relation expressed by IO, etc.), and (2) the animacy values of the arguments. Probabilistic functions embodying the statistical regularities sketched above are needed to map these features onto primary versus secondary placement options. Length and definiteness may

add some further refinements. A generator incorporating such an algorithm is currently under development at our institutes.

Finally, we should point out that the approach taken here cannot uncover linear order determinants related to the sentence's information structure and its discourse context (e.g. contrastiveness). Such factors may modulate the above preferences or license "tertiary" linear orders beyond the primary and secondary options considered here.

References

1. Langkilde, I., Knight, K.: Generation that exploits corpus–based statistical knowledge. In: Proceedings of the 36th ACL & 17th COLING, Montreal (1998)
2. Wasow, T.: Postverbal behavior. CSLI Publications, Stanford CA (2002)
3. Müller, G.: Optimality, markedness, and word order in German. Linguistics **37** (1999) 777–815
4. Keller, F.: Gradience in grammar: Experimental and computational aspects of degrees of grammaticality. Unpublished Ph.D. thesis, Univ. of Edinburgh (2000)
5. Skut, W., Krenn, B., Brants, T., Uszkoreit, H.: An annotation scheme for free word order languages. In: Proceedings of the Fifth ANLP, Washington D.C. (1997)
6. Uszkoreit, H.: Word Order and Constituent Structure in German. CSLI Publication, Stanford CA (1987)
7. Pechmann, T., Uszkoreit, H., Engelkamp, J., Zerbst, D.: Wortstellung im deutschen Mittelfeld. Linguistische Theorie und psycholinguistische Evidenz. In: Perspektiven der Kognitiven Linguistik. Westdeutscher Verlag, Wiesbaden (1996)
8. Kurz, D.: A statistical account on word order variation in German. In Abeillé, A., Brants, T., Uszkoreit, H., eds.: Proceedings of the COLING Workshop on Linguistically Interpreted Corpora, Luxembourg (2000)
9. Kempen, G., Harbusch, K.: A corpus study into word order variation in German subordinate clauses: Animacy affects linearization independently of grammatical function assignment. In Pechmann, T., Habel, C., eds.: Multidisciplinary approaches to language production. Mouton De Gruyter, Berlin (in press)
10. Hawkins, J.A.: A performance theory of order and constituency. Cambridge University Press, Cambridge (1994)
11. Abbott, B.: Definiteness and indefiniteness. In Horn, L.R., Ward, G., eds.: Handbook of Pragmatics. Blackwell, Oxford (in press)
12. König, E., Lezius, W.: A description language for syntactically annotated corpora. In: Proceedings of the 18th COLING, Saarbrücken (2000)
13. Kempen, G., Harbusch, K.: How flexible is constituent order in the midfield of German subordinate clauses? A corpus study revealing unexpected rigidity. In: Proceedings of the International Conference on Linguistic Evidence, Tübingen (2004)