Toward a uniform psycholinguistic framework for the linearization of nominal and verbal constituents during sentence production: Evidence from Dutch and German treebanks

Gerard Kempen, Max Planck Institute for Psycholinguistics, Nijmegen
Karin Harbusch, Department of Computer Science, University of Koblenz-Landau

Experimental-psycholinguistic and corpus-linguistic studies of linearization in sentence production have focused on the ordering of nominal constituents (NPs). This research has uncovered a variety of linearization determinants. Well-known are the following: (1) Topic << Comment, Theme << Rheme, Given << New, Definite << Indefinite; (2) Actor << Undergoer, Subject << Object; (3) Pronominal << Full; (4) Animate << Inanimate; (5) Focus << Background; (6) Light << Heavy.

A uniform account of these factors is possible based on the following postulates:
— Sentence production comprises a Conceptualization stage followed by a Grammatical Encoding stage (Kempen, 1977).
— These stages have the option of delivering partial structures incrementally (Kempen & Hoenkamp, 1982, 1987; Levelt, 1989).
— Grammatical Encoder assigns a ready-to-be-linearized NP to the leftmost free grammatically allowed position on the right of the most recently linearized constituent (Kempen & Harbusch, 2003, 2004; Kempen, 2013).

In such a system, highly available conceptual fragments tend to be delivered to the Grammatical Encoder at early points in time (e.g. given vs. new referents; animate vs. inanimate entities), and to be assigned to early positions in the evolving syntactic structure (Bock & Warren, 1985). This also holds for syntactic fragments that take up less time to be assembled by the Grammatical Encoder (e.g., pronominal NPs as compared to full NPs; light vs. heavy NPs).

In the present paper, we investigate whether this theoretical framework is applicable to verbal constituents as well, i.e., to clauses. We assume that every lexical, auxiliary, modal or copula verb functions as Head of exactly one finite or nonfinite clause. As for the time course of clause planning during speaking, we assume that Main clauses tend to be grammatically encoded prior to Subordinate clauses. In Dutch and German, the obligatory verb position depends on clause type: Main clauses are “Verb-second” (SVO; V2); Subordinate clauses are “Verb-final” (SOV; Vf). Hence, during spontaneous speaking, Main clauses put heavier demands on early availability of Head verbs than Subordinate clause—on two counts: Main << Subordinate, and V2 << Vf. On the further assumption that corpus frequency can be viewed as an index of the online availability (conceptual, lexical, or otherwise) of the verb at production time, we predict that high-frequent verbs can meet the early availability demands more readily than low-frequent verbs.

As part of an extensive ongoing corpus study of the incidence of SVO and SOV orders in Dutch and German, we extracted all occurrences (tokens) of finite and nonfinite verb forms in two spoken and two written treebanks (Table 1; Kempen & Harbusch 2012, 2013). In each treebank, we determined the total corpus frequency of each inflected verb form (type). The tokens were categorized as (A) Finite-Main, i.e. V2; (B) Finite-Subordinate, i.e. Vf; or (C) Nonfinite-Subordinate, also Vf.

The chief result so far is strong confirmation of our prediction. For each verb form (type), the proportion of its tokens in category A (Finite-Main), relative to its tokens in the other categories, rises with the total corpus frequency of that verb form (see the rising curve in Figure 1). Even considering the set of finite verb forms alone (categories A and B), we observe that high-frequent forms are more likely to belong to category A (V2) relative to category B (Vf) than low-frequency forms (Table 2). Furthermore, the order Finite-Main << Finite-Subordinate clause turns out to be much more frequent than the inverse order (about 75% vs. 25% in both spoken treebanks).

In conclusion, the frequency patterns support the notion of early verb availability as facilitating early verb position. This, in turn, corroborates the hypothesis that the same temporal mechanism underlies the linearization of nominal and verbal constituents during spontaneous spoken language production.
Table 1. The four treebanks and their size in terms of number of verb form tokens.

<table>
<thead>
<tr>
<th>Treebank</th>
<th>Approximate total number of verb form tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIGER (written/DE) (Brants et al., 2004)</td>
<td>106,000</td>
</tr>
<tr>
<td>Verbmobil (spoken/DE) (Stegmann et al., 2007)</td>
<td>50,000</td>
</tr>
<tr>
<td>Alpino (written/NL) (van der Beek et al., 2002)</td>
<td>18,000</td>
</tr>
<tr>
<td>CGN 2.0 (spoken/NL) (van Eerten, 2007)</td>
<td>168,000</td>
</tr>
</tbody>
</table>

Table 2. Proportions of finite verb form tokens in Main and Subordinate clauses, in function of their corpus frequency. The number of verb form types is mentioned in the columns labeled N.

<table>
<thead>
<tr>
<th>Corpus frequency of verb form</th>
<th>CGN (Dutch)</th>
<th>Verbmobil (German)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>Main</td>
</tr>
<tr>
<td>High-frequent</td>
<td>118</td>
<td>.50</td>
</tr>
<tr>
<td>Low-frequent</td>
<td>1267</td>
<td>.42</td>
</tr>
</tbody>
</table>

Figure 1. Proportion of verb form tokens in Main clauses (category A) relative to tokens in Subordinate clauses (categories B and C; spoken treebanks only). Abscissa: token frequency of verb form types expressed in the form of percentile ranges. Ordinate: proportion of verb forms in Main clauses. NB Trends in written treebanks are similar but less clear-cut—presumably due to post-editing.

References