Funded by:

DFG

and **Karst-Stiftung** (University of Konstanz)

Universität Konstanz

COMPARATIVE GERMANIC SYNTAX WORKSHOP 34
CGSW 34

University of Konstanz
June 14-15, 2019

Main Organizers:
Andreas Trotzke
George Walkden

Organizing Board:
Josef Bayer
Miriam Butt
Nicole Dehé
Tanja Kupisch
Theo Marinis
Andreas Trotzke
George Walkden

Information-structural effects of the accessibility of finite verbs: Corpus evidence from spoken English, Dutch, and German

Gerard Kempen[1] and Karin Harbusch[2]
[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]Faculty of Computer Science, University of Koblenz-Landau, Koblenz, Germany

Speakers of Germanic languages tend to PRIORITIZE sentence constituents whose content and form was planned with little processing effort: "Easy" constituents are more likely assigned to early (anterior) positions than constituents that were harder to plan, provided the grammar allows sufficient linear-order flexibility. Well-known factors promoting anterior placement are CONCEPTUAL ACCESSIBILITY (including TOPIC-COMMENT and ANIMACY) and LEXICAL ACCESSIBILITY (FREQUENCY, PRIMING). Empirical work on these information-structural effects tend to focus (pro)nominal constituents—presumably because nouns and pronouns enjoy much intraclausal freedom of position. In the present contribution, we show that finite verbs, too, are subject to effects of accessibility, in spite of very restricted placement options.

The work to be reported here began as fallout from a corpus study on a different topic (Kempen & Harbusch 2017). As part of that study, we had counted frequencies of verbs functioning as head of a main clause (always finite: "Main-Fnt"), of a finite subordinate clause ("Sub-Fnt"), or of a non-finite subordinate clause or VP ("Non-Fnt"). Below, we refer to this variable as "ClauseType". The data had been extracted from three corpora consisting of syntactically annotated spoken sentences extemporaneously and unscriptedly produced by native speakers of German (VERBMOBIL), Dutch (CORPUS GESPROKEN NEDERLANDS), and English (SWITCHBOARD; for corpus details and references, see Kempen & Harbusch 2017). Prior to the present project, we had lemmatized the inflected verbforms and calculated the total frequency ("TotFreq") of each verb lemma by adding its Main-Fnt, Sub-Fnt and Non-Fnt occurrences. Pairs of homonymous or polysemous lemmas had been treated as a single lemma (e.g., the TotFreq of Eng. *be* includes all its occurrences, whether as auxiliary, copula or main verb). The result is a list of 7531 unique verb lemmas (German: 1083; Dutch: 3884; English: 2564).

We ranked the lemmas by TotFreq and calculated, for each lemma, its proportion of Main-Fnt, Sub-Fnt, and Non-Fnt tokens (see Fig.1). This revealed a remarkable statistical interaction between TotFreq and ClauseType: Whereas the Sub-Fnt proportions are more or less constant—as expected *a priori*—, the Main-Fnt proportions appear to rise with increasing TotFreq of the lemmas. Stated differently, high-frequent verbs tend to be OVER-REPRESENTED IN MAIN CLAUSES, whereas their presence in finite subordinate clauses tends to align with their TotFreq. (By implication, low-frequent verbs are over-represented in non-finite clauses; the proportion of non-finite tokens equals 1 minus the sum of the finite proportions; not shown in Fig.1)

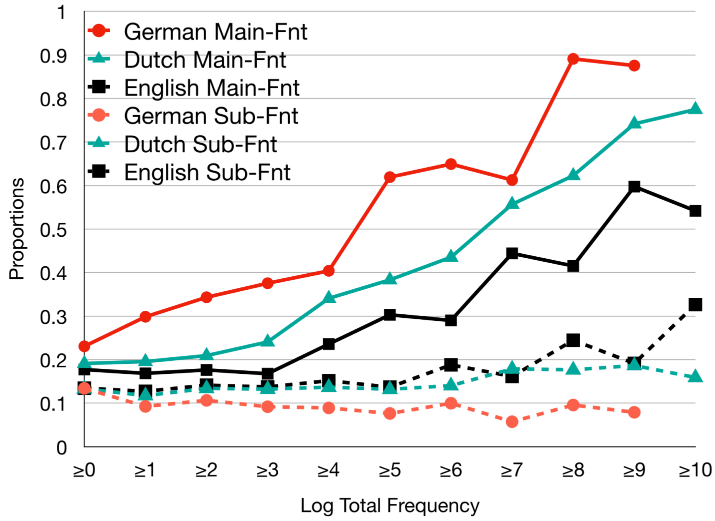

Figure 1. Main-Fnt proportions (continuous curves) and Sub-Fnt proportions (dashed curves) as a function of Log_e(TotFreq), ClauseType, and Language. The proportions are "unweighted": Every verb, however high its TotFreq, has the same influence on the calculated mean proportions.

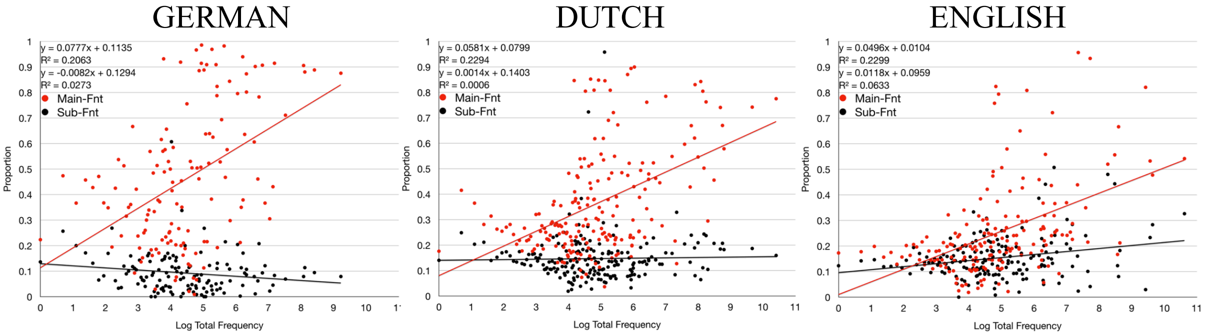

Figure 2. Effect of verb accessibility on MCB and ClauseType in the three target languages. Every red dot in a scatter diagram represents the mean Main-Fnt proportion of a GROUP of verbs that all have the same TotFreq in their corpus; likewise, every black dot shows the Sub-Fnt proportion of such a group. For instance, the average Main-Fnt and Sub-Fnt proportions of all hapax verbs (TotFreq =1) are depicted above the Log(TotFreq)=0 value (i.e., in the leftmost pair of a red and a black dot in each diagram); and the rightmost pairs of dots show the proportions of a single verb (*sein/zijn/be*). The trendline equations are shown in the top-left corners.

We refer to the rise of the Main-Fnt proportions with increasing TotFreq as MAIN-CLAUSE BIAS OF HIGH-FREQUENT VERBS—for short: “MCB effect”. The size of the MCB effects in each of the target languages is indexed by the combined differences (1) between the slopes of the linear trendlines (red minus black), and (2) between their intercepts (black minus red). Figure 2 reveals substantial cross-language differences regarding the MCB effect sizes: German > Dutch > English. Statistical analysis using Beta Regression confirmed significance of the TotFreq * ClauseType * Language interaction.

In our account of the two phenomena we focus on the contrast between German and Dutch on one hand, and English on the other—considering that the German vs. Dutch difference may very well be an artifact due to the smaller corpus size of, and the narrower range of topics addressed in, the German corpus compared to the Dutch one (the Dutch and English corpora are more similar in these respects).

We propose to account for the existence of the MCB effect, and for its cross-language size differences, in terms of a direct and an indirect processing consequence of verb accessibility. The DIRECT one follows from the ACCESSIBILITY–ANTERIORITY LINK that also explains other information-structural phenomena. High-frequent, hence rapidly accessible finite verbs are more likely to present themselves as candidate fillers for anterior clause positions than lower-frequent verbs. This benefits V2 clauses (German/Dutch Main-Fnts) more than V3 clauses (English Main/Sub-Fnts); and Vfinal clauses are unlikely to benefit at all (German/Dutch Sub-Fnts). The early availability of a suitable finite verb may have an INDIRECT consequence on the main vs. subordinate status of the clause-under-construction. If the proposition underlying a clause has autonomous illocutionary force (assertional, interrogative) it usually can be delivered not only as a main but also as a subordinate clause (cf. non-restrictive relative clauses). In such cases, early availability of the finite verb creates a favorable occasion for the clause to be realized in the form of a main clause, thus boosting the percentage of main clauses in the corpora, especially of main clauses with high-frequent verbs. The data pattern we obtained for the three languages fits in with this explanation.

In support of the first part of our hypothesis (the accessibility-anteriority link), we refer to related corpus work we did using published treebanks for ancestral varieties of the target languages (Old High German, Old Saxon, and Old English). There, finite verbs have a wider range of placement options than in the present-day varieties. Indeed, in OHG, OS and OE main clauses, high-frequent verbs tend to occupy earlier positions than low-frequent verbs (Harbusch, Kempen & van Kemenade, 2019). Currently, we are exploring how, in strict SOV languages such as Japanese and Korean, accessibility-anteriority effects on verbs get blocked.