Verb frequency, clause typing, and verb placement in language production: A corpus study with treebanks for spoken and written English, Dutch, and German

Gerard Kempen (MPI Nijmegen) & Karin Harbusch (Informatics, Uni-Koblenz)

In recent work using treebanks of spoken German, Dutch, and English, we compared the frequency distributions of finite verbs in main clauses and finite subordinate clauses (henceforth "subclauses") (K&H 2019, LCN journal). Although largely overlapping, the distributions within main clauses revealed an upward shift relative to the distributions in subclauses—in all three languages (Fig. 1). We called this effect "Main-Clause Bias of high-frequency verbs" (MCB). We also observed a remarkable cross-language difference: the MCB is significantly stronger in German and Dutch than in English, as indicated by steeper trendlines (Fig. 2). We linked these findings to the default positions of the finite verb in main vs. subclauses. In Dutch and German, these positions differ more strongly (main: verb-second; sub: verb-final) than in English (verb-third in main and subclauses, except for some special main-clause constructions featuring "residual verb second"). We reasoned that high-frequency verbs are easier and more rapidly accessible to the clause planning system; therefore, if the clause-under-construction needs an anterior verb position (e.g., verb-second), high-frequency verbs are better suited to fulfill this need than low-frequency ones, thus supporting a fluent utterance continuation.

Since the above publication, we have explored *written*-language treebanks for the same three languages. They revealed substantially reduced but still significant effect sizes for the MCBs in German and Dutch, whereas the already small MCB for spoken English has disappeared (Fig. 2, right panel). The demand for rapid accessibility of anterior verbs is lower during writing than during speaking. The remaining cross-modal differences between MCB effect sizes suggests that faster accessibility is not the entire story.

A clue to an additional causal factor emerged from a comparison of the number of main clauses (#main) and the number of subclauses (#sub) in each treebank. We observed a substantial correlation of the #main:#sub ratio in a corpus with the size of its MCB effect (Fig. 3; Spearman Rho = .77. N = 6, $p < .05$). (The differing ratios are not visible in the charts of Fig. 1, which are based on *normalized* frequencies.) This correlation means that clause-typing decisions (selection of main vs. subordinate format) are not entirely at the mercy of unchangeable syntactic/pragmatic/lexical choices made earlier, but are also affected by "preferences" of currently active verbs for an early or a late clause position.

How to implement this notion of preference? We start from the—presumably uncontroversial—assumption that, in the three target languages, clauses specify two possible target positions for the finite head: one *unmarked/default* (verb-final in Dutch and German, verb-third in English), one *marked* (verb-second in Dutch/German, residual verb-second in English). To this, we add the assumption that verbs differ w.r.t their value on a continuous *marked–unmarked scale*: a low "positional markedness" value means that the verb has a strong basic tendency to occupy the posterior clause position, meaning that some cognitive effort is needed to coerce it into the anterior position. A high value means that anterior placement requires little or no special effort. The current markedness value of a given verb presumably depends not only on accessibility but also on other factors (e.g., how often it is used in a clause expressing assertive or interrogative illocutionary force).

Given this theory, we can account (descriptively if not explanatorily) for the data pattern we found. (1) The MCB effect is due to the fact that high-markedness verbs react faster and more reliably to the demand for anterior verb placement in main clauses than low-markedness ones. The markedness parameter plays no role in subclauses, where finite verbs standardly select the posterior position. (2) Time/fluency pressure weighs less heavily in written than in spoken language. (3) The small #main:#sub ratio for English compared to Dutch and German is attributable to the high level of similarity between the English main and subclause formats compared to the rather sharp contrast in Dutch and German (see above). The similarity between the English clause formats implies that the utility of learning and applying associations between verbs and positional-markedness values carries much less weight than it does in German and Dutch.

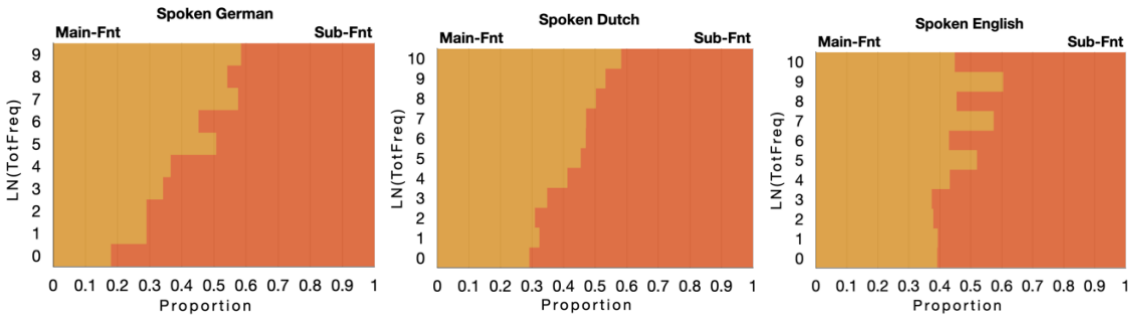

Fig. 1. Normalized proportions of finite main and finite subordinate clauses at various levels of Total verb lemma Frequency. In each corpus, the proportions of Main-Finite and Sub-Finite verbs in a given frequency class add up to 1. LN(TotFreq) = natural logarithm of a verb's total frequency, which includes all finite and non-finite occurrences.

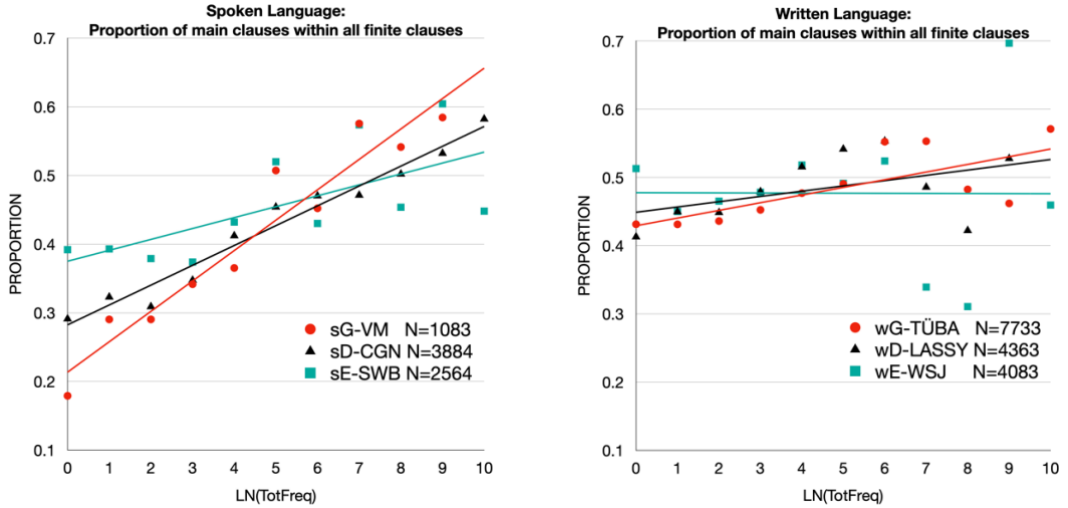

Fig. 2. Left panel: another view of the spoken language data in Fig. 1, now with linear trendlines. Right panel: Normalized proportions for written language, corresponding to the proportions depicted in the left panel. Abbreviations: s = spoken; w = written; G/D/E=German/Dutch/English; VM = VerbMobil corpus; CGN = Corpus Gesproken Nederlands (Spoken Dutch); SWB = SwitchBoard corpus; TÜBA = TüBa-D/Z (German newspaper "taz"); LASSY = LASSY Small (LArge Scale SYntactic Annotation of Written Dutch); WSJ = Wall Street Journal corpus; N = the number of different verb lemmas per corpus.

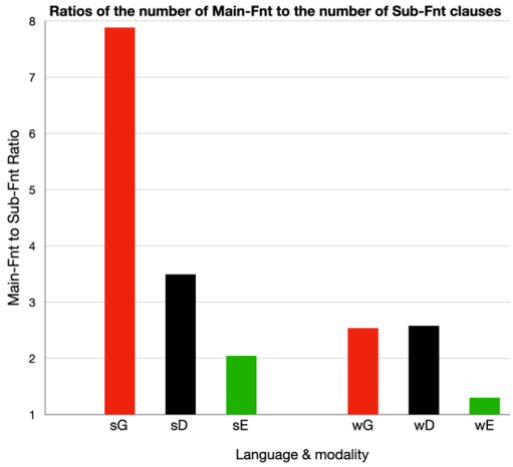

Fig. 3. Ratios of the total number of Main-Fnt clauses to the total number of Sub-Fnt clauses. (The number of finite clauses equals the number of finite verbs; we disregard clauses that underwent finite-verb ellipsis.)