

A treebank study of clausal coordinate ellipsis in spoken and written language

Karin Harbusch¹ & Gerard Kempen^{2,3}

¹Department of Computer Science, University of Koblenz-Landau

²Max Planck Institute for Psycholinguistics, Nijmegen

³Cognitive Psychology Unit, Leiden University

One of the benefits of incremental sentence production is reduction of the working memory capacity needed for advance planning: The planning units can be considerably smaller (measured in terms of word length) than in case of non-incremental production. The same advantage has been claimed for the various forms of ellipsis, which preempt the need to plan the detailed shape of one or more constituents and thereby reduce the size of planning units. Because working memory load tends to be higher in spoken than in written language, one expects that speakers, in comparison with writers, will more frequently resort to the use of elliptical constructions. However, in two corpus studies into the incidence of Clausal Coordinate Ellipsis (CCE) in spoken and written English, Meyer (1995) and Greenbaum & Nelson (1999) obtained a data pattern opposite to this prediction: In written clausal coordinations, the proportion of CCE versions was about twice as high as in spoken coordinations. The pattern was explained in terms of audience design: Non-elliptical (unreduced) clauses include more repetition and thereby facilitate comprehension.

Recent treebanks with large numbers of hand-parsed spoken (CGN2.0) and written (ALPINO) Dutch sentences, enabled us to verify the data pattern for another language: In written Dutch, the percentage of elliptical versions within the set of all clausal coordinations was even three times higher than in spoken Dutch: 34% versus 11% (Table 1).

However, a breakdown of the CCE sentences into various forms of CCE revealed a phenomenon that cannot be explained in terms of audience design. Two important forms of CCE distinguished in the linguistic literature are Forward Conjunction Reduction (FCR; as in (1)) and Gapping (2). (Two less frequent forms, not discussed any further here, are SGF (3) and RNR (4).) In FCR, each conjunct has its own explicit head verb; in Gapping, the second conjunct has no head verb. In each of the treebanks, Gapping and FCR together covered 92% of the CCE cases (with the remaining 8% more or less evenly distributed among SGF and RNR). However, the distribution of FCR and Gapping in the two treebanks differed widely. Whereas in written clausal coordinations Gapping accounted for only 10% of the CCE cases (with a large majority of 82% embodying FCR), in spoken clausal coordinations the incidence of Gapping was much higher: 31% (leaving 61% for FCR).

We attribute both phenomena—both the higher proportion of elliptical versions within the set of all clausal coordinations and the shifted frequencies of CCE forms—to a narrower scope (“window”) of online grammatical planning in spoken as compared to written sentence production. More specifically, in order not to overtax online working memory load, speakers have a stronger tendency than writers to plan the grammatical shape of each clause in isolation, that is, without taking the shape of coordinated clauses into account. As a consequence, they overlook many elliptical options. That this tendency reduces the incidence of FCR follows from the linguistically plausible assumption that clauses are planned as projections of verbs. In FCR, the second (elliptical) conjunct has its own overt head verb and therefore is planned as a new clause. In Gapping, however, the second clause is planned not as the projection of a verb but rather as a modification or extension of an existing (the first, non-elliptical) clause, much like a substitution repair (Kempen, 2009). Hence, because Gapping involves one overt verb and one clause only, it is less likely to overburden online working memory.

In sum, the data suggest (1) that CCE in spontaneous speech benefits the speaker, not the listener; and (2) that Gapping should be analyzed as a monoclausal structure-with-revisions rather than as a partly deleted biclausal structure.

Table 1. Clausal coordination and CCE in the Dutch treebanks

Treebank	Average sentence length	Percentage of sentences with a clausal coordination	Percentage of clausal coordinations with CCE
ALPINO (written)	17.8	13	34
CGN 2.0 (spoken)	8.6	6	11

- (1) FCR a. Last year, Emma lived in Nijmegen and ... worked in Amsterdam
 b. The town [S where Paula works and ... Harry lives]
- (2) Gapping a. You live in Nijmegen and your son ... in Amsterdam
 b. Conrad commutes to Leiden, and ... usually by train
- (3) SGF Into the woods went the hunter and ... shot a hare
- (4) RNR Simone submitted one ... and Agnes reviewed two abstracts