

Incremental sentence production and clausal coordinate ellipsis: A treebank study comparing spoken and written language in Dutch and German

Karin Harbusch

University of Koblenz-Landau
Computer Science Department
Universitätsstr. 1
56070 Koblenz
Germany

HARBUSCH@UNI-KOBLENZ.DE

Editors: David Schlangen and Hannes Rieser

Abstract

From two corpus studies into varieties of clausal coordination in English (Meyer, 1995 and Greenbaum & Nelson, 1999), it is known that the incidence of clausal coordinate ellipsis (CCE) is about two times higher in written than in spoken language. We present a treebank study into CCE in written and spoken Dutch and German which confirms this tendency. Moreover, we observe considerable differences between written and spoken language with respect to the incidence of four main types of clausal coordinate ellipsis—*Gapping*, *Forward Conjunction Reduction (FCR)*, *Backward Conjunction Reduction (BCR)*, and *Subject Gap with Finite/Fronted Verb (SGF)*. We argue that the detailed data pattern cannot be accounted for in terms of audience design, and propose an explanation based on the assumption that during spontaneous speaking—but not during writing—the scope of online grammatical planning is basically restricted to one (finite) clause.

Keywords: Treebanks, Human Language Production, Incrementality, Coordination, Coordinate Ellipsis, Gapping, Conjunction Reduction, Audience Design

1 Introduction

One of the benefits of INCREMENTAL SENTENCE PRODUCTION is a reduction of the working memory capacity needed for advance planning: The planning units can be of considerably smaller size (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 both to incremental production and to the use of elliptical constructions. As far as we know, this prediction is generally borne out for incremental production. However, in two corpus studies into the incidence of clausal coordinate ellipsis in spoken and written English, Meyer (1995) and Greenbaum & Nelson (1999) obtained a data pattern opposite to the prediction. In written clausal coordinations, the proportion of elliptical versions is about twice as high as in spoken clausal coordinations.

Recent treebanks with spoken and written sentences for Dutch and German enable us to test whether the latter finding generalizes to other Germanic languages, and to explore how spoken and written language production affect the incidence of four different types of clausal coordinate ellipsis—GAPPING, FORWARD CONJUNCTION REDUCTION (FCR), BACKWARD CONJUNCTION REDUCTION (BCR), and SUBJECT GAP WITH FINITE/FRONTED VERB (SGF). As far as we know, the only treebank study of CCE was carried out by Zinsmeister (2006). In a

quantitative study with the German TAZ¹ Treebank, she identified 8,133 sentences (37% of the total number of sentences) with a coordination of syntactic constituents of ANY type. She only reports one number dealing with CCE types: 83 sentences embodying SGF, i.e., about 1 percent. (This proportion is not directly comparable to the one we report below because we only take into account sentences with CLAUSAL coordinations.)

In this paper, we explore treebanks for spoken and written Dutch and German. For Dutch, we probe the ALPINO treebank (van der Beek et al., 2002) that consists of about 7,000 sentences of newspaper text, and CGN2.0 (van Eerten, 2007) with about 130,000 spoken sentences (or rather dialogue turns) from more than ten different domains. For German, we explore the TIGER treebank (Brants et al., 2004) of written newspaper text with about 50,000 sentences, and the spoken utterances in the VERBMOBIL dialogues in TüBa-D/S (about 38,000 sentences).

We first introduce the four types of clausal coordinate ellipsis (Section 2). In Section 3, we present the detailed results of the four treebank studies—first the Dutch study (Section 3.1), then the German study (Section 3.2), and finally some striking between-modality differences and cross-language similarities (Section 3.3). In Section 4, we propose a theoretical explanation of the data, focusing on the differential effects that incremental sentence production exerts on the individual CCE types. Finally, in Section 5, we sum up and mention desiderata for future work.

2 A typology of clausal coordinate ellipsis

In the linguistic literature on clausal coordination one often distinguishes four main types of coordinate ellipsis² (for overviews, see van Oirsow, 1987; Steedman, 2000; Sag et al. 2003; te Velde, 2006; and Kempen, 2009):

- GAPPING, with three variants called:
 - LONG DISTANCE GAPPING (LDG),
 - SUBGAPPING, and
 - STRIPPING,
- FORWARD CONJUNCTION REDUCTION (FCR),
- BACKWARD CONJUNCTION REDUCTION (BCR; also known as *Right Node Raising* or *RNR*), and
- SUBJECT GAP WITH FINITE/FRONTED VERB (SGF).

Table 1 illustrates these CCE types in terms of examples taken from the VERBMOBIL dialogues in TüBa-D/S and CGN2.0.³ We adopt the following PSYCHOLINGUISTICALLY MOTIVATED definitions of CCE types. They derive from work by Kempen (2009) who argues that coordinations are structurally similar to self-repairs in spontaneous speech, i.e. can be viewed as special type of “update” constructions (cf. Section 4.1). With respect to the underlying linguistic framework, we try to be as theory-neutral as possible. However, we presuppose a separation between the HIERARCHICAL and the LINEAR STRUCTURE of sentences. In descriptions of linear order, we use the terminology of topological fields (with Forefield, Midfield, and Endfield as translations of VORFELD, MITTELFELD, and NACHFELD, respectively; cf. Höhle, 1986). As the encodings of hierarchical structures in the four treebanks differ

¹ TüBa-D/Z treebank (Hinrichs et al. 2004) is a corpus of German newspaper texts currently comprising about 22,000 sentences taken from the Wissenschafts-CD of “die tageszeitung” (TAZ). Henceforth, it is called the “TAZ” treebank in order to avoid confusion with TüBa-D/S (Stegmann et al., 2000), i.e. a spoken German treebank for the VERBMOBIL domain (Wahlster, 2000), which we will call the “VERBMOBIL” treebank in the following.

² We do not deal here with the elliptical constructions known as VP Ellipsis, VP Anaphora and Pseudogapping because they involve the generation of pro-forms instead of, or in addition to, the ellipsis proper. For example, *John laughed, and Mary did, too*—a case of VP Ellipsis—, includes the pro-form *did*. Nor do we account for re-casts of clausal coordination as coordinate NPs (e.g., changing *John likes skating and Peter likes skiing* into *John and Peter like skating and skiing, respectively*). Presumably, such conversions involve a semantic rather than syntactic mechanism.

³ We were unable to identify tokens of Long-Distance Gapping (LDG) in the VERBMOBIL corpus. In CGN2.0, we found nine exemplars. The theory proposed in Section 4 may explain why these numbers are so small in spoken text.

GAPPING (g)	(1) <i>Nachdem ich selbst ungern in die Oper gehe und</i> As I myself reluctantly to the opera go and nachdem_g Sie so gerne in die Oper gehen_g ... you so readily ‘As I myself go to the opera reluctantly and you [do so] readily ...’
LDG (g) ⁺ g	(2) <i>Hij had zich verkleed als meisje en</i> he had himself dressed-up as girl and <i>z'n vriend had_g [zich verkleed]_{gg} als oude vrouw</i> his friend as old woman ‘He had dressed up as a girl and his friend [had dressed up] as an old woman’
SUBGAPPING (sg)	(3) <i>Dann können wir uns erst um die Verbindung kümmern und</i> Then can we ourselves first for the connection care and <i>dann [können wir]_{sg} die Hotels aussuchen</i> then the hotels select ‘First we may take care of the connection and afterwards [we may] select the hotels’
STRIPPING (str)	(4) <i>Am Freitag hätte ich bis elf Uhr Zeit bzw.</i> On Friday would-have I until eleven o'clock time and [am Freitag hätte ich]_{str} ab dreizehn Uhr auch wieder Zeit_{str} onward-from 1PM o'clock also again ‘On Friday, I would have time until 11AM and from 1PM onward too’
FCR ∅	(5) <i>Wenn Sie schon in dem Parkhotel waren und</i> If you already in the Park-hotel stayed and [wenn Sie]_f das gut fanden, ... it ok found ‘If you already stayed in the Park hotel and found it OK’ (6) <i>Da wäre Hotel X, [welches am Bahnhof liegt und</i> There would-be hotel X which at-the station lies and welches_f zum Zentrum 15 Minuten Laufzeit hat to-the center 15 minutes walking-time has ‘There is hotel X which is located at the station and [which is] a 15 minutes walk to the center’
BCR (b)	(7) <i>Im Juli hätte ich nur einen [Tag frei]_b und</i> In-the July would-have I only one and <i>im Mai zwei Tage frei</i> in-the May two days off ‘In July, I only have one, and in May two days off’
SGF (s)	(8) <i>Dann fahren wir frühmorgens los und</i> Then get we early-morning started and [wir]_s sind mittags da are at-noon there ‘Then, we get started early morning and [we'll] arrive at noon’

Table 1. CCE examples in spoken German and Dutch. Struck-out text represents elisions.

substantially—e.g., not all treebanks use VP nodes (cf. Section 3)—we suppose the tree structures to be rather flat.

All forms of GAPPING (cf. examples (1) to (4) in Table 1) are characterized by elision of the posterior member of a pair of lemma-identical Verbs.⁴ The position of this Verb need not be peripheral but is often medial, as in (2) through (4). Every non-elided constituent

⁴ In our definitions of CCE types, we restrict ourselves to coordinations encompassing two conjuncts, called ANTERIOR (first, left) and POSTERIOR (second, right), respectively (cf. Footnote 13).

(“REMNANT”) in the posterior conjunct should pair up with a constituent in the anterior conjunct that has the same grammatical function but is not coreferential.⁵ Stated differently, the members of such a pair are CONTRASTIVE—in (1), the Subjects *ich selbst* ‘I myself’ vs. *Sie* ‘you’, and the Modifiers *ungerne* ‘reluctantly/dislike-to’ vs. *gerne* ‘readily/like-to’. Only contrastive constituents are expressed overtly in the posterior conjunct. This restriction rules out cases as *John eats apples and Peter ~~eats~~ in the car*. Notice that *gehen* ‘go (3rd Person)’ in the posterior conjunct of (1) can be elided although it has no wordform-identical (but only a lemma-identical) anterior counterpart (wordform *gehe* in the first conjunct is 1st Person, Singular, Present Tense of the Verb *gehen* ‘go’).

Gapped sentences resemble answers to implicit multiple *wh*-question. For instance, Steedman (1990:248) writes: “[E]ven the most basic gapped sentence like *Fred ate bread, and Harry, bananas* is only really felicitous in contexts which support (or accommodate) the presupposition that the topic under discussion is *Who ate what?*”. According to Reich (2007), Gapping answers an implicit⁶ *wh*-question. Hartmann (2000) examines the close connection between focus and ellipsis in Gapping. This correspondence is reflected in the particular intonation contour typical of Gapping structures.

In LONG DISTANCE GAPPING (LDG), the remnants originate from different clauses (more precisely: from different clauses that belong to the same SUPERCLAUSE; a superclause is a hierarchy of finite or nonfinite clauses that—with the possible exception of the topmost clause—do not include a Subordinating Conjunction. In (2), *hij* ‘he’ belongs to the main clause headed by the Verb *had* ‘had’ but *zich* ‘himself’ and *als meisje* ‘as a girl’ to the nonfinite complement clause headed by the Past Participle *verkleed* ‘dressed up’.

In SUBGAPPING, the posterior conjunct includes a remnant in the form of a nonfinite complement clause (VP; *aussuchen* ‘select’ in (3)). In STRIPPING, the posterior conjunct is left with one non-Verb remnant, often supplemented by a sentential Adverb such as *auch* ‘too’ (in example (4), *auch wieder* ‘again’), or a negation. Notice that our definition of Gapping in terms of the hierarchical structure of the conjuncts circumvents word order issues which are not relevant for Gapping. For instance, (9) exemplifies one alternative to the word order in example (4). Actually, each of the constituents *Zeit*, *ich*, *am Freitag*, and *auch wieder* can go to the Forefield.

- (9) ... *ab* *13 Uhr* *hätte* *ich am Freitag auch wieder Zeit*
 from-onward 13 o'clock would-have I on Friday also again time

In STRIPPING, there is only one remnant, i.e. contrastive constituent. However, we also count as Stripping those cases where the posterior conjunct introduces new semantic aspects of the event described in the anterior conjunct (see example (10) from the VERBMOBIL corpus).

- (10) *Ich möchte gerne nach Hannover, und zwar über Karneval nächstes Jahr*
 I would like(to-go) to Hannover and namely during carnival next year
 ‘I want to go to Hannover, (and) namely during carnival next year.’

In FORWARD CONJUNCTION REDUCTION (FCR), elision affects the posterior token of a pair of left-peripheral strings consisting of one or more wordform-identical major constitu-

⁵ We distinguish three identity relationships between constituents in coordinated conjuncts: lemma identity, wordform identity and coreferentiality. For LEMMA IDENTITY, only the lexical entries (CITATION FORM) of the constituents have to be identical; WORDFORM IDENTITY requires, in addition, identity of their morphological features. COREFERENTIAL CONSTITUENTS refer to the same discourse entity or entities, irrespective of whether or not they include the same lemma(s). Lemmas are “syntactic words”; their lexical entries specify the sentential environments in which they are allowed to occur. The morphological and phonological information associated with words is specified in another type of lexical entry called word forms or lexemes.

⁶ Example (i) below might be analysed as implicit question answering based on Gapping, more precisely, on Stripping. Here, the speaker formulates a question. The elliptical construction contains the tentative answer—now by the same speaker. (In this example, *mit dem Flugzeug* can hardly have been intended as an extraposed part of the question, assuming that in the VERBMOBIL context *wie* queries the means of transportation.

(i) *Wie sollen wir uns dahin begeben, mit dem Flugzeug?*
 How shall we ourselves there move with the plane
 ‘How shall we go there, by plane?’

ents. In (5), the posterior tokens of *wenn Sie* ‘if you’ and in (6), *welches* ‘which’, respectively, belong to such a pair and are eligible for FCR. Notice that if the Finite Verb belongs to the left-peripheral string, FCR and Gapping are not always distinguishable. We count such constructions as FCR. Example (11) is a case in point.

- (11) *Dann können wir noch Essen gehen oder*
 Then can we even eat go or
~~*dann können wir noch*~~ *irgendwas in der Richtung tun*
 something in the direction do
 ‘Then we can go dining or do something like that’

BACKWARD CONJUNCTION REDUCTION (BCR) is almost the mirror image of FCR as it deletes the anterior member of a pair of right-peripheral lemma-identical word strings (*Tag(e)* ‘day(s)’ in (7)); however, BCR may elide PART OF a major constituent—e.g. only the part *Tag(e)* of the Direct Object in (7). In addition, it requires only lemma identity (cf. Number Singular vs. Plural of the elided Noun in example (7)).⁷ BCR does not necessarily elide the Verb (see (12)).

- (12) *Erst hören wir uns und dann sehen wir uns wohl am Bahnhof*
 First hear we ourselves and then see we ourselves probably at-the station
 ‘First we phone each other, and then we see each other at the station’

SUBJECT GAP WITH FINITE/FRONTED VERB (SGF; see Wunderlich, 1988, and Höhle (1983) who calls the phenomenon “SLF Koordinationen”, for “Subjektlucken in finiten/frontalen Sätzen”; for a recent survey, see Kathol, 2001) elides the Subject of the posterior conjunct in a main clause, when in the anterior conjunct the wordform-identical Subject follows the Finite Verb (Subject-Verb inversion). Elision of the posterior Subject cannot be due to FCR since the anterior Subject is not left-peripheral. Furthermore, the initial constituent of an anterior SGF conjunct should NOT be an argument. This is illustrated by the ill-formed ellipsis in example (13) where a Complement clause opens the anterior conjunct. In SGF case (8), the initial constituent is an Adjunct.

- (13) **Das Examen bestehen will er und ~~er~~, kann auch*
 The exam to-pass wants he and can too
 ‘He wants to pass the exam and is also able to [do this]’

We also subsume under the heading of SGF cases like (14), where the anterior conjunct is a conditional subordinate clause rather than a main clause. (See Höhle (1990) and Reich (2008) for discussion of the affinity between this structure and SGF.)⁸

- (14) ... *dann reicht es ja, wenn wir ungefähr um neun losfahren würden und*
 then suffices it already if we about at nine get-started_{inf} would and
~~*wir*~~ *würden dann mittags dort ankommen*
 would then at-noon there arrive
 ‘... then it suffices already if we would leave at nine and we would arrive there at noon’

3 CCE in spoken and written Dutch and German

In this section, we present a detailed overview of the incidence of the four types of clausal coordinate ellipsis in Dutch and German treebanks with spoken and written text. We first describe our study for Dutch with the Treebanks ALPINO for written newspaper text, and CGN2.0 for spoken text. Then, we report the findings obtained with the TIGER treebank of

⁷ Notice that example (7) cannot be analyzed as a case of one-anaphora, for at least two reasons. First, the anaphoric element is in the anterior rather than the posterior conjunct. Second, anaphoric use requires *eins* instead of *ein*, as in example (ia/b). (*Tag* has Masculine, *Auto* Neuter Gender.)

(ia) *Ich habe ein und du zwei Autos.*

‘I have one and you two cars’

(ib) *Ich habe zwei Autos und du eins/*ein*

‘I have two cars and you one’

⁸ Notice that the word order in the second conjunct of example (5) distinguishes FCR from SGF: In SGF, the second conjunct always has main-clause word order (i.e. yielding *und fanden das gut* ‘and found it OK’ instead of *und das gut fanden* in (5)).

the German written newspaper text, and with the TüBa-D/S treebank that contains the spoken VERBMOBIL dialogues.

The linguistic encodings in the four treebanks under consideration are rather different. We cannot describe the individual formats in detail for reasons of space (but see the Appendices for relevant specifications of clausal coordination in the four treebank formats).

In order to obtain comparable numbers, we defined search patterns that implement the definitions of the four CCE types and take into account differences between the specific notations used in the four treebanks (such as whether or not they encode VPs and VP coordination). All retrieved sentences were manually classified with respect to CCE type in a uniform manner. If one coordination embodies more than one CCE type, as in (15), (17) or (18) below, we counted them as separate instances. The incidence of a CCE type in a corpus is expressed as the number of exemplars of that CCE type in the total number of sentences with at least one coordination.

3.1 CCE in Dutch

Here, we report the results of our comparative study of CCE in written and spoken language in Dutch (also see Harbusch & Kempen, 2009a).

3.1.1 CCE in the ALPINO treebank

The ALPINO Treebank, released in November 2002, contains 7,153 syntactically annotated Dutch sentences. All annotations in this corpus were manually inspected. The sentences comprise the full (cdbl newspaper) part of the Eindhoven corpus (Uit den Boogaart, 1975).

Figure 1 illustrates the encoding format for CCE in the ALPINO treebank. In sentence (15), the Subject (labeled *su*) and the Verb Complement (labeled *vc*) are elided due to FCR and BCR, respectively. In ALPINO, elisions are encoded by COREFERENTIAL INDICES⁹ at the remnant and its corresponding empty leaf node. For instance, the index 1 is attached to the overt Subject of the anterior conjunct, and to the Subject of the posterior conjunct, where the Subject has been deleted due to FCR. Index 2 denotes the Verb Complement, which has been erased from the anterior conjunct due to BCR. (Notice that, in the hierarchical structure, the Verb Complement has been encoded as part of the first conjunct although it has been elided due to BCR and occurs overtly as remnant in the second conjunct.) For more details on the labels used in the Figure, see Appendix A.

(15) ***Het moet en zal een Nederlands stuk worden***

It should and will a Dutch piece become

‘It should become and it will become a Dutch piece’

We identified 931 sentences with at least one clausal coordination. Within this set, we found 319 CCE cases of any type. The detailed proportions of different CCE types are reported in Section 3.1.3, where they are compared to those in CGN2.0.

3.1.2 CCE in the CGN2.0 treebank

CGN stands for *Corpus Gesproken Nederlands* ‘Corpus of spoken Dutch.’ The speech data originate from adult speakers of standard Dutch in Flanders (one third of the corpus material) and the Netherlands (two third). Version CGN2.0, released in 2004, comprises 130,594 sentences from various dialogue domains. The treebank has been encoded in the NEGRA format and can be inspected using the TIGERSearch tool (König & Lezius, 2003). SECONDARY EDGES, represented by curved edges in tree diagrams such as Figures 2, relate remnants to the root node where the constituent is supposed to have been elided. In Figure 2, the secondary edge expresses that the Head (cf. edge label HD) *bent* ‘handed’ is supposed to be a child of the SSUB node as well. However, the target position within the list of siblings is intentionally left undetermined. (For more details on the encodings used in CGN2.0, see Appendix B.) In

⁹ The COREFERENTIAL INDEX also encodes Subject-to-Subject Raising, indicating that the Subject of the infinitival Verb Complement is identical with the Subject of the governing Finite Verb. We ignored all non-elliptical uses of these indices.

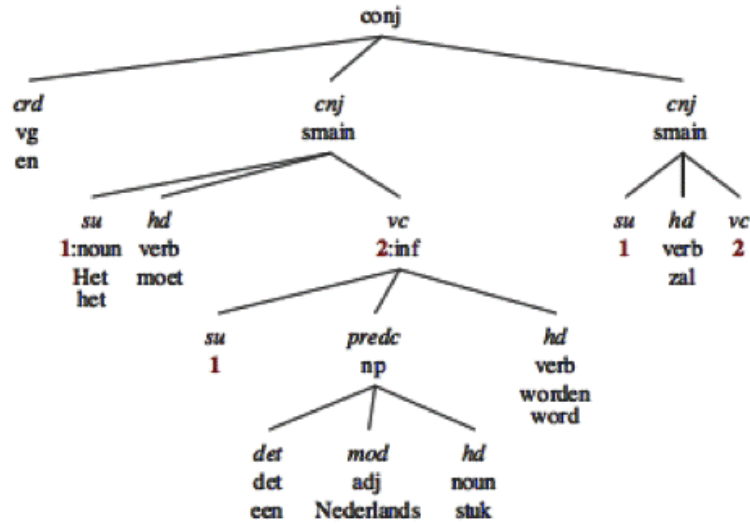


Figure 1. Hierarchical structure of sentence (15) in ALPINO.

example (16)—part of a longer utterance—the last word *bent* ‘are’ in the second conjunct has been deleted from the first conjunct due to Backward Conjunction Reduction (BCR). We found 8,653 sentences with at least one clausal coordination. Within this set, we counted 924 CCE instances.

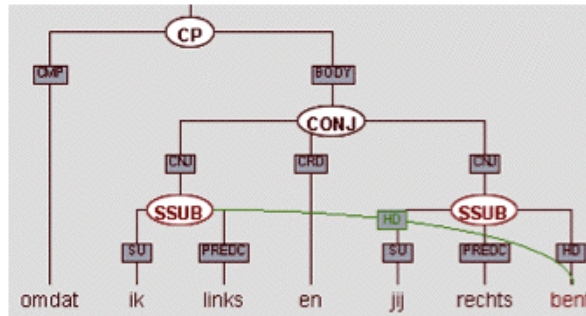


Figure 2. Tree structure of example (16) in CGN2.0.

- (16) ... *omdat ik links en jij rechts bent* ...
 because I left-handed and you right-handed are
 ‘because I am left-handed and you [are] right-handed’

3.1.3 CCE incidence in spoken and written Dutch

In written Dutch, the percentage of elliptical versions within the set of sentences with at least one clausal coordination is three times higher than in spoken Dutch: 34% versus 11% (see Table 2). This means that the lower incidence of CCE in written than in spoken language that has been observed in English, holds for Dutch as well.

Treebank	Average sentence length (words)	Total number of sentences with clausal coordination	Total number of CCE	Percentage of sentences with clausal coordination	CCE percentage
ALPINO (written)	17.8	931	319	13	34
CGN 2.0 (spoken)	8.6	8,653	924	6	11

Table 2. Clausal coordinate ellipsis in the Dutch treebanks. The numbers in the rightmost column are percentages of the number of sentences containing one or more clausal coordination.

3.2 CCE in German

In this section, we test whether the differing CCE tendencies in spoken and written language generalize to German (also see Harbusch & Kempen, 2009b).

3.2.1 CCE in the TIGER treebank

The TIGER Treebank, released in December 2003, contains 50,474 German syntactically annotated sentences from a German newspaper corpus. As illustrated in Figure 3¹⁰, TIGER’s annotation scheme uses many clause-level grammatical functions (Subject, Direct and Indirect Object, etc., shown as edge labels in the sentence diagrams). As already mentioned for CGN2.0, elided constituents in coordinate clauses are represented by secondary edges (represented by curved arrows pointing from the remnant to the root of the elided element). Like in our CGN2.0 exploration, we deployed the TIGERSearch tool to design queries that retrieve clausal coordination (whether elliptical or not). We classified the elliptical ones (those including one or more secondary edges) into one of the four CCE types by hand. (For details, see Harbusch & Kempen, 2007, 2009c.) We found 7,194 sentences with at least one clausal coordination. Within this set, we counted 4,020 CCE instances.

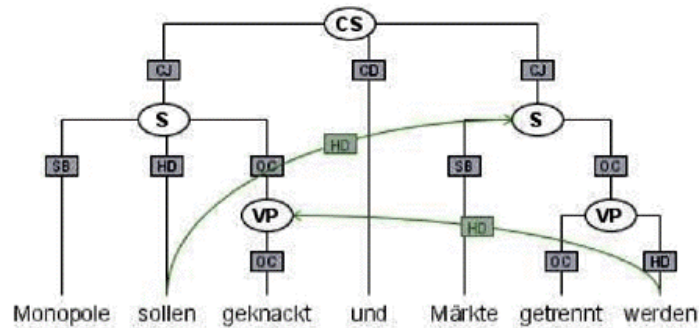


Figure 3. Tree diagram for example (17): Subgapping combined with BCR. The two remnants of the posterior clause are *Märkte* ‘markets’ and the VP (nonfinite clause) *getrennt werden* ‘be separated’.

- (17) *Monopole sollen geknackt und Märkte getrennt werden*
 Monopolies should shattered and markets split be
 ‘Monopolies should be shattered and markets should be split’

3.2.2 CCE in the VERBMOBIL treebank

The VERBMOBIL treebank is encoded in the same manner as the TAZ corpus (Hinrichs et al., 2004) but rather differently from TIGER (see Lemnitzer and Zinsmeister, 2006, page 82, for a comparison of the tag sets; cf. Appendix D).

The VERBMOBIL treebank comprises 38,328 utterances in fifteen different subcorpora. We selected and retrieved all trees via TIGERSearch. However, the VERBMOBIL treebank lacks annotations that relate remnants and elisions (cf. the indices in ALPINO, and the secondary edges in CGN2.0 and TIGER). Therefore, we classified all coordinated clauses for CCE type manually. See Appendix D for encoding details.

Figure 4 illustrates the structural encoding of example (18). This sentence exhibits “forward” elision of *viertel vor zwölf könnte* combined with “backward” elision of *abholen*.

- (18) *Viertel vor zwölf könnte ich Sie oder mein Fahrer Sie abholen*
 Quarter to twelve could I you or my driver you up-pick
 ‘Quarter to twelve, I could pick you up or my chauffeur could do so’

¹⁰ The trees in Figures 2 and 3—from CGN2.0 and TIGER, respectively—look similar because they both originate from the TIGERSearch tool. However, the grammatical annotations are not identical. For instance, CGN2.0 does not use VP nodes. A combination of automatic sentence retrieval and manual inspection of the hits enabled us to obtain comparable numbers (see Appendix C for encoding details).

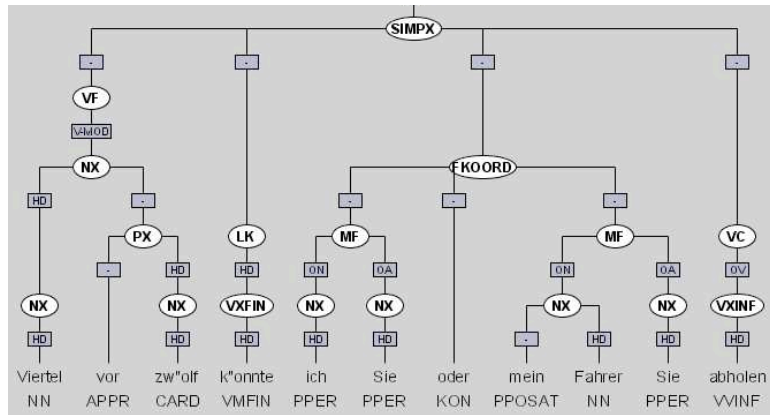


Figure 4. Encoding of example (18) in VERBMOBIL in terms of a coordination (node label FCOORD) in the Midfield (MF) of the Subjects (grammatical function ON in squared boxes) and the Direct Objects (OA) of the conjuncts.

This example also illustrates a particularity of spoken language, namely UNDERREDUCTION with respect to backward elision of *Sie* ‘you’ from the first conjunct. (Without this *Sie*, we would analyze the example as a coordination of NPs rather than as clausal coordination—see Section 4.3)

In total, we found 3,713 VERBMOBIL sentences with at least one clausal coordination. This set includes 1,314 CCE tokens.

3.2.3 CCE incidence in spoken and written German

In written German, the percentage of elliptical versions within the set of clausal coordinations is considerably higher than in spoken German: 56% versus 35%. This means that the lower CCE incidence in spoken than in written text holds not only for English and Dutch but also for German.

Treebank	Average sentence length (words)	Total number of sentences with clausal coordination	Total number of CCE	Percentage of sentences with clausal coordination	CCE percentage
TIGER (written)	17.6	7,194	4,020	14	56
VERBMOBIL (spoken)	9.9	3,713	1,314	10	35

Table 3. Clausal coordinate ellipsis in the German treebanks. The numbers in the rightmost column are percentages of the number of sentences containing one or more clausal coordination.

3.3 Incidence of CCE types in spoken versus written language

Table 4 shows the relative frequencies in German and Dutch of the four CCE types we distinguish. It reveals striking differences between modalities and similarities across languages. In both Dutch and German, FCR and Gapping together take the lion’s share of CCE instances: 80 to 92 percent. However the distribution of FCR and Gapping differs considerably between the spoken and written modalities. In spoken language, Gapping is responsible for about one third of the CCE cases (32%), in written language for only one eighth of the CCE cases (13%). In both languages, the relative frequencies of the CCE types in the spoken vs. the written production modality differ significantly from one another: $\chi^2 = 192.3$, $df = 3$, $p < 0.0001$ for German; $\chi^2 = 66.1$, $df = 3$, $p < 0.0001$ for Dutch. In the German treebanks, BCR¹¹ and

¹¹ As BCR is defined as elision of the right periphery of the anterior conjunct, and both Dutch and German often have verbs in clause-final position, one may wonder how often BCR causes elision of verbs (and/or separable verb

CCE type	Spoken language		Written language	
	VERBMOBIL (German)	CGN 2.0 (Dutch)	TIGER (German)	ALPINO (Dutch)
Gapping	33%	31%	17%	10%
FCR	55%	61%	63%	82%
BCR	1%	3%	10%	5%
SGF	11%	5%	10%	3%

Table 4. Relative frequencies of the four types of CCE, expressed as percentages of the total set of sentences exhibiting CCE.

SGF are well represented (in particular SGF) whereas in the Dutch corpora they live a somewhat marginal existence.

In the next section, we propose an explanation for the results of our corpus study, focusing on the two main aspects in which spoken language appears to differ from written language: (1) a lower incidence of CCE, and (2) within the CCE types a higher incidence Gapping.

4. Incremental sentence production and CCE

As argued in the introduction, given the usually tighter processing constraints during speaking than during writing, one expects speakers, in comparison to writers, to resort more frequently to the use of elliptical constructions. However, the corpus work by Meyer (1995) and Greenbaum & Nelson (1999) as well as the present treebank study yield a data pattern opposite to this prediction. In written clausal coordination, the proportion of elliptical versions is about twice as high as in spoken coordination.

A tentative explanation of these data appeals to AUDIENCE DESIGN (cf. Bell, 1984): Speakers try to reduce the cognitive load imposed on their dialogue partners—here, by avoiding ellipsis. Greenbaum & Nelson (1999), following Meyer (1995), propose a similar explanation on page 116:

“Repetition helps the listener to understand what is being said by making the discourse less dense. Full forms, which involve repetition, tend therefore to be preferred over elliptical forms in speech.”

However, the empirical evidence for audience design as a systematic speaker strategy aiming to preclude comprehension problems is mixed (e.g., Arnold, et al., 2004; Haywood et al., 2005; Brennan & Hanna, 2009). Moreover, this hypothesis cannot explain our observation that, in both languages, the incidence of Gapping is higher in the spoken than in the written modality.¹²

In the following, we present another explanation for the complete data pattern.¹³ It is in line with incremental sentence production (see Section 4.1 for a brief introduction). Given the

particles, which are always clause-final), and whether the frequencies of such elisions vary between modalities. We checked the corpora for such cases but they are rare and do not exhibit any salient differences.

¹² An additional argument against an explanation in terms of audience design is suggested by the strong tendency to use telegraphic style in VERBMOBIL. See examples (i) and (ii), where the hearer has to figure out which constituent(s) is/are left out. In both examples, the clause-initial position of both conjuncts is empty: ‘topic drop’.

(i) *startet Bonn Hauptbahnhof um acht Uhr fünfundvierzig und*
 leaves Bonn main-station at 8 hour 45 and
kommt an in Hannover Hauptbahnhof um zwölf Uhr vier
 arrives in Hannover main-station at 12 hour 4
 ‘(It) leaves Bonn main station at 8:45 AM and arrives at the main station of Hannover at 12:04PM’
 (ii) *liegt zentral, hat Hallenbad und Fitneßraum*
 is located centrally has indoor-pool and fitness-room
 ‘(It) is located centrally, has indoor-pool and fitness-room’

¹³ Although in this paper we focus on binary coordinations (i.e. with two conjuncts), the explanation to be presented below will generalize to n -ary coordinations on the assumption that these are produced in a pairwise manner. That is, all pairs of conjoined (super)clauses (1,2), (2,3), ..., ($n-1,n$) are checked for applicability of a CCE operation. E.g., an n -ary BCR construction suppresses the right-periphery in all but the last clause. See, e.g., Borsley (2005) for additional restrictions on n -ary coordinations.

stronger time pressure during speaking, we assume that, in order not to overtax current workspace capacity, speakers have a stronger tendency than writers to plan the grammatical shape of each clause IN ISOLATION. Consequently, speakers are less likely to take into account the shape of coordinated clauses than writers, who have more opportunities for looking ahead/back. Hence, CCE constructions whose well-formedness does depend on both conjunct's grammatical shape (i.e. FCR, BCR and SGF) are expected to occur less frequently in spoken language. In contrast, Gapping's relative frequency will actually INCREASE because in Gapping only ONE clause needs to be considered at planning time (in a repair-like manner; cf. Section 4.2). Moreover, we present in Section 4.3 more evidence from the VERBMOBIL corpus for the limitation of the workspace capacity. Finally, we sum up the line of argumentation in Section 4.4.

4.1 Key features of incremental sentence production and Kempen's (2009) psycholinguistically motivated ellipsis generation model

According to Levelt's (1989) widely accepted "blueprint of the speaker", sentence production is a five-stage process:

- INTENDING: creating the communicative intention,
- CONCEPTUALIZING: mapping the intention onto a set of concepts and conceptual (thematic) relations; output from this stage is a conceptual structure ("message") in the form of a tree with branches whose linear order is undefined,
- GRAMMATICAL ENCODING: mapping concepts onto lemmas; output is the hierarchical/dominance and the linear/precedence structure of the sentence,
- PHONOLOGICAL ENCODING: taking the linearly ordered tree as input, the spoken form of the sentence is produced by replacing the lemmas by lexemes (phonological word-forms), and
- ARTICULATING (not further addressed here).

All mapping information during these stages stems from the MENTAL LEXICON, in particular concepts, conceptual (thematic) relations, and lemmas. (For present purposes, we simply assume a one-to-one correspondence between concepts and lemmas.) Lemmas will be attached to syntactic trees as terminal nodes. Associated with every lemma is information specifying how it can be used in sentences. For instance, a lemma that corresponds to an action concept specifies how the concept's thematic relations are mapped onto grammatical functions (e.g., Actor onto Subject and Patient/Theme onto Direct Object). Lexemes (inflected or uninflected wordforms) are yet another type of lexical entries in the Mental Lexicon.

The creation of a communicative intention during speaking and writing is not always an "indiscernible event" yielding precisely the full meaning underlying the next sentence at one point in time. Often it is a "time-extended event" during which the meaning is conceived gradually, fragment by fragment. This allows the language user to initiate grammatical encoding before the complete communicative intention is available—for instance already when the "topic" has been conceived but not yet the "comment". Through this INCREMENTAL mode of sentence production, language users reduce the risk of overtaxing the currently available processing resources, in particular the WORKSPACE for the formation and short-term maintenance of grammatical structures. Of course, this strategy increases the risk of grammatically incoherent utterances; but these can be repaired online, in particular when they threaten to become incomprehensible.

For psycholinguistic evidence regarding incremental sentence production see, e.g., Kempen & Hoenkamp, 1982, 1987, and Levelt, 1989. For grammar formalisms that take incrementality into account, e.g. Tree Adjoining Grammar, see Ferreira 2000 and Frank & Badecker, 2001. For Dynamic Syntax, see Purver & Kempson, 2004; for a formalism-independent parsing method based on unsupervised-learning from Treebanks, see, Seginer, 2007.

As mentioned in Section 2, our account of CCE phenomena is inspired by Kempen's (2009) theory of coordinate ellipsis in Dutch and German. The four types of clausal coordinate ellipsis—Gapping, FCR, BCR and SGF—are argued to originate in four different stages

of sentence production. Moreover, he classifies the phenomena in terms of appropriateness repair-likeness or unlikeness.

REPAIRS occur when—halfway through, or at the end of, a sentence—speakers modify the communicative intention underlying the current utterance in such a way that at least part of the utterance needs to be UPDATED. In such repairs, some or all of the originally intended content that already has been encoded conceptually and grammatically and surfaced as an overt utterance, is replaced by more appropriate content, which requires at least a partially different overt realization.

The ellipsis phenomena FCR and Gapping can be classified as a repair process according to Kempen. In case of Gapping, the conceptual content underlying the Verb is shared between conjuncts. This also holds for the thematic relations contracted by the Verb and for the mappings between thematic relations and grammatical functions. Only some Arguments or Adjuncts need to be replaced or added. Consequently, Kempen states that Gapping comes into existence already in the Conceptualization phase as an updating operation. In contrast, FCR takes place during Grammatical Encoding, basically because there the Verb concepts do not get updated. (For the detailed argumentation, see Kempen, 2009.)

Kempen (2009:679) argues that BCR actually is not a form of coordinate ellipsis:¹⁴

“Coordinate structures only afford a suitable playing ground for Left Deletion as they often give rise to contrastive pairs. The plausibility of viewing BCR as a form of coordinate ellipsis [...] is extremely low anyway: the notion of updating entails FORWARD ellipsis only because, by definition, the update comes later than the original structure.”

Finally, Kempen states that posterior clauses of SGF coordinations do not “borrow” a Subject NP in the course of an incremental updating operation. Instead, in the communicative intention underlying an SGF coordination, the speaker assigns several predicates to the referent of the subject NP SIMULTANEOUSLY.

In the following, we assume that in spontaneous speech language production is more often incremental than in writing due to higher time pressure and more severe workspace limitations.

4.2 “One-Clause” versus “Multiple-Clause” CCE

As shown in Section 3, the relative frequency of Gapping is higher in spoken language than in written text whereas the other CCE types have lower relative frequencies in spoken than in written language. Why?

Steedman (2000:182) claims that Gapping constructions do NOT result from elision, and thus, there is no need to analyze sequences of non-clausal constituents (in the posterior conjunct) as headless “clauses”. We follow this argument, which is in line with Kempen (2009). We assume that Gapping does NOT involve constructing a posterior clause but that the posterior conjunct is built by modifying the anterior one. The whole process of Gapping¹⁵ works much like speech REPAIR (as outlined in Section 4.1) where only the corrections, i.e. the contrastive elements, are uttered. Hence, only ONE clause—more precisely: one SUPERclause¹⁶—

¹⁴ It is known (see, e.g. Hudson, 1976) that BCR-like structures arise also in non-coordination contexts (cf. example (i), where the relative clauses modifying Subject and Direct Object allow ellipsis similar to the pattern in the coordinated NPs in example (ii), from the TIGER corpus). We have not searched for such cases in spontaneous speech.

(i) Politicians who fought for ~~chimpanzee rights~~ may well snub those who have fought against *chimpanzee rights*

(ii) ... das richtige Gleichgewicht zwischen und
the right balance between and
die Gleichseitigkeit von *verschiedenen ... Allokationsmechanismen*.
The equilaterality of different allocation mechanism

¹⁵ Gapping in coordinate structures is similar to Gapping in question answering. Producing a gapped answer presupposes that the structure built up during parsing the question is reused in the answer generation process (see, e.g., Branigan et al., 2000).

¹⁶ According to our definition, Gapping in Dutch and German requires some syntactic checking, namely the inspection of superclause boundaries (cf. Section 2). Thus, literal translations into Dutch and German of example (i), taken from Culicover & Jackendoff (2005, p.273), do not work in the intended meaning because the Subordinating Conjunction blocks any wider scope of Gapping.

needs to be kept in the workspace simultaneously, i.e. we claim that Gapping is a “ONE-CLAUSE” CCE construction. Consequently, all constituents needed to construct two clausal Gapping conjuncts can be kept in the workspace at the same time. In contrast, in written text, writers are more likely to resort to NP coordinations for expressing the same information. Stripping, in particular, often becomes recast in this manner, presumably for stylistic reasons.

In contrast, applying FCR, BCR or SGF presupposes a grammatical processing window spanning more than one clause (“MULTIPLE-CLAUSE” CCE construction). Application of FCR requires comparing the left-hand peripheries of two or more conjoined clauses. Likewise, BCR demands comparisons between the right-hand peripheries of the clausal conjuncts. In SGF, the encoding mechanism needs to verify the positions and the referential identity of the Subject constituents of consecutive clauses. The “one-clause” property of Gapping entails that Gapping-type CCE structures can survive the limitations due to limited workspace capacity more easily than structures embodying other CCE types.

Finally, we comment on the fact that BCR seems “anti”-incremental. BCR requires the second conjunct to be grammatically planned at least to some extent in parallel with the first one, because the potential elision in the anterior conjunct is licensed by the right-peripheral part of the posterior conjunct. This constraint limits incrementality in spontaneous speech, i.e. the utterance of the final anterior conjunct relies on the posterior conjunct’s syntactic shape. When starting our project, we expected BCR to be absent from spoken corpora. This expectation is even strengthened by alternative ellipsis options, which would affect the second conjunct, imposing even fewer constraints on incremental processing, and erase the same amount of text. For instance, in example (18) above, Gapping would elide *viertel vor zwölf, könnte, Sie* and *abholen* (cf. sentence (25) in Section 4.3, which is actually shorter). The same holds for example (17), of which example (19) is the Gapping variant.

- (19) *Monopole sollen geknackt werden und Märkte getrennt*
 Monopolies should shattered be and markets split
 ‘Monopolies should be shattered and markets split’

But, why would speakers resort to BCR in a small percentage of the corpus material? There should be another explanation for BCR constructions. Often the two conjunct are very short (cf. (20) or (12) above) so that the two conjuncts may simultaneously fit into the workspace. Alternatively, the shared right-peripheral constituent(s) might have been added as a kind of afterthought to both already uttered constituents (cf. (21)). However, this claim needs additional evidence from psycholinguistic studies into human CCE production under controlled circumstances.

- (20) ... *wann Ebbe **ist** und wann Flut **ist***
 when low-tide and when high-tide is
 ‘... when is ebb and flow’
- (21) ... *wir wollten uns für ein Meeting treffen bzw. eines ausmachen **in Hannover, wenn***
 we wanted us for a meeting meet and-resp. one negotiate in Hannover when
ich mich recht entsinne
 I myself correctly remember
 ‘...we wanted to have a meeting or negotiate one in Hannover if I remember correctly’

4.3 Additional evidence for limited workspace capacity in speaking

In the VERBMOBIL treebank of spontaneous spoken utterances, we found many dialogue turns that are UNDERREDUCED¹⁷, i.e., where ellipsis options were not used, or used only par-

(i) *Robin believes that everyone pays attention to you when you speak French, and Leslie, German*

As stated in Footnote 3, we could not find any token of Long-Distance Gapping in VERBMOBIL (which might have a domain-specific reason: identical subjects in both conjuncts prevail in that corpus) and only nine tokens in CGN2.0. This observation may be viewed as evidence for the claim that having several (super)clauses in the workspace at the same time easily leads to overtaxing the currently available capacity, and to avoidance of LDG structures.

¹⁷ UNDERREDUCTION also occurs in written text. However, there it is used intentionally and suits stylistic purposes such as emphasizing the non-reduced constituent—cf. (i) from the TIGER corpus.

tially (see example (22)). For instance, in example (18) above (repeated here as (23)), the *Sie* ‘you’ could be left out resulting in an NP coordination (as in (24)) or in Stripping/Gapping (as in (25)), which is alternatively interpretable as including a discontinuous NP). Such incomplete elisions attest to the possibility of workspace overload when two conjuncts need to be planned simultaneously.

(22) *Ich **organisiere** die Flüge und Sie **organisieren** also dann das Hotel*

I organize the flights and you organize well then the hotel’

‘I organize the flights and you organize then the hotel’

(23) *Viertel vor zwölf könnte ich **Sie** oder mein Fahrer **Sie** abholen*

Quarter to twelve could I you or my driver you up-pick

‘Quarter to twelve, I could pick you up or my chauffeur could do so’

(24) *Viertel vor zwölf könnte ich oder mein Fahrer Sie abholen*

Quarter to twelve could I or my chauffeur you up-pick

‘Quarter to twelve, my chauffeur or me could pick you up’

(25) *‘Viertel vor zwölf könnte ich Sie abholen oder mein Fahrer*

Quarter to twelve, could I you up-pick or my chauffeur

In support of workspace limitations as a causal factor, we can also point to typical errors in VERBMOBIL. Here, inaccurate structural matches between a remnant in a posterior conjunct and its counterpart in the anterior conjunct occur frequently (cf. FCR of *das/da* in example (26)). These might also be interpreted as consequences of workspace overload.

Moreover, telegraphic style and CCE production often go together, as in example (27) where Gapping has been applied to the incomplete first conjunct.

(26) ***das** ist direkt am Hauptbahnhof und ~~da~~ kostet das Einzelzimmer*

that is directly at-the main-station and costs the single-room

ehundert und neunundzwanzig Mark.

one-hundred and twenty-nine Mark

‘that is directly at the main station and (there) a single room is 129 DM’

(27) ***hat** sogar Schwimmbad **dabei** und ~~hat~~ Bar **dabei***

has even swimming-pool included and bar included

‘(it) has included even swimming pool and bar’

In view of these observations, there seems to be a tendency in incremental sentence production to erase from the workspace some or all constituents of an already uttered clause as soon as the Conceptualizer to the Grammatical Encoder embark on a new (super)clause.

4.4 The reasoning in a nutshell

- A. The five stages of language production are subject to constraints on processing resources, in particular on the workspace for Conceptualization and Grammatical Encoding of Communicative Intentions.
- B. The workspace capacity available to the language user will, on average, be smaller in speaking than in writing.
- C. The smaller the currently available workspace capacity, the higher the likelihood of incremental sentence production.
- D. When the grammatical encoding mechanism is forced to operate with a limited workspace capacity, it has few opportunities to plan multiple (super)clauses at the same time.
- E. During speaking, clauses are more often grammatically encoded in isolation than during writing—as a consequence of B and D.
- F. Gapping is a “One-Clause” CCE structure, whereas FCR, BCR or SGF are “Multiple-Clause” CCE structures.
- G. Gapping can survive the limitations due to workspace limitations more easily than structures embodying other CCE types—due to E and F.

(i) *Der Lehrkurs bei ihm war streng und war gründlich*

The lesson by him was strict and was intensive

‘His lesson was stern and intensive’

- H. Hence, the trend towards more Gapping in spoken than in written language production.
Q.E.D.

5 Conclusion

Summing up, we have presented data verifying, for Dutch and German, the data pattern emerging from two corpus studies into the incidence of Clausal Coordinate Ellipsis in spoken and written English. We proposed a new explanation for the relative frequencies of the four CCE types we distinguish, based on incremental sentence production and the more limited workspace capacity in speaking than in writing.

So far, we have avoided any excursions into incremental PARSING as it would lead too far away from the central topic addressed here and would be highly speculative. At least in Computational Linguistics, parsing elliptical constructions is a difficult problem (see, e.g., Kübler et al., 2009), partially due to the fact that neither of the conjuncts might consist of a complete, grammatically correct clause (cf. (17)).

How does a human hearer understand CCE? Let us assume that parsing is highly parallel (in line with new studies on human parsing (cf. Boston et al, in press), and that sentence parsing is the inversion of sentence generation (cf. bidirectional grammars such as Optimality Theory; Prince & Smolensky, 1993/2004). Due to frequency, the hearer is most likely prepared for the Gapping variant of CCE. Thus, the perception of a Coordinate Conjunction triggers the hypothesis (among a set of parallel hypotheses) that the current workspace content (i.e. the anterior conjunct) will be modified by repair-like substitutes. Accordingly, the hearer tries to find a contrastive constituent for every incrementally constructed constituent of the second conjunct. This strategy is doomed to failure as soon as a Verb is parsed. Consequently, FCR and SGF parsing hypotheses for the second conjunct get activated. However, these CCE hypotheses compete with the more highly frequent ellipsis-free hypotheses, which may allow the workspace to be emptied completely. FCR and SGF parsing means, that a prefix (a left-peripheral part) of the first conjunct is assumed to be the valid prefix of the second conjunct as well. (N.B. for SGF, the length of the prefix is exactly one constituent, and by definition it has to be the Subject; for FCR, the parser might try a language-specific default instead of trying all prefixes systematically.) Accordingly, the workspace can only be emptied partially before the second conjunct is parsed.

BCR is triggered by a parsing failure (or a misparse¹⁸) of the first conjunct. Thus, at the very end of the second conjunct, a re-parse (or a reordering of parallel parsing hypotheses) of the first conjunct is called for. Consequently, the working memory should NOT be emptied before the complete clausal coordination has been parsed. We hasten to that these considerations are highly speculative—even more so because sometimes combinations of CCE types (such as example (17)) need to be parsed.

Gapping benefits from incremental processing particularly, as only the contrastive constituents of the second conjunct have to be checked, and can be uttered in any order. Detailed frequency studies into comparatives such as *John is prouder of his dog than Mary is proud of her cat* (see, e.g., Lechner, 2004), which are reminiscent of Gapping, and arguably similar, are needed.

Acknowledgement

I would like to express my deep gratitude to Gerard Kempen, who co-authored several related articles on the present topic. I also thank the two anonymous reviewers for their valuable comments. However, none of them can be blamed for remaining errors—any misconceptions and shortcomings in this paper are my own.

¹⁸ Separable Verb prefixes are often elided due to BCR, which changes the meaning considerably (e.g. replacing *rufen* ‘shout’ by *anrufen* ‘call up’ as in *Ich rufe und du schreibst Peter an* ‘I call up Peter and you write to him’. Here, the first conjunct could have been parsed as ‘I shout’.

Appendices

In the following, we list the nomenclature for coordinate structures in the four different tree-banks investigated in the present study. As illustrated by the examples in Section 3, the encodings are rather different. Therefore, we had to abstract away from the details of the tree-bank formats. However, due to space restrictions, we cannot enumerate the various search patterns for the corpora.

Appendix A: ALPINO

In ALPINO, a node labeled “conj” dominates nodes labeled conjunct (*cnj*) and coordinator (*crd*), respectively. Edges are not labeled but the node labels denote not only a grammatical function (e.g., *su* for Subject) but, in addition, either a part-of-speech (e.g., *vg* for the Coordinating Conjunction such as *en* ‘and’ in Dutch) supplemented with the word in the sentence, or a phrasal category. In order to license clausal coordination, the node labels of conjuncts should be either *smain* for Main Clause or *ssub* for Subordinate Clause. Within a clause, grammatical functions are labeled: Subject as *su*, the Finite Verb as *hd* and the Verb Complement as *vc*, etc.

Elisions are encoded by COREFERENTIAL INDICES at nodes. The remnant and its corresponding empty leaf node get identical numerical indices. Which of the two is expanded can only be identified in the sentence itself.

We wrote a JAVA program to automatically retrieve all clausal coordinations with and without such indices, and manually assigned CCE types to the coordinations containing an index. The detailed proportions of different CCE types are reported in Section 3.1.3.

Appendix B: CGN2.0

In CGN2.0, edges are labeled with the grammatical functions. Edge labels are represented as squared boxed at the edges in tree diagrams provided by TIGERSearch (see, e.g., Figure 2 in Section 3.1.2). Nodes carry either lexical information or a phrasal category. In coordination, a node labeled “CONJ” dominates edges labeled with *CNJ* (conjunct) and *CRD* (coordinator). In order to represent clausal coordination, *CNJ*-edges end at *SMAIN* or *SSUB* nodes, denoting main and subordinate clauses, respectively. Within a clause, grammatical functions are represented at edges. For instance, *SU* = Subject, *PREC* = PREDiCate, and *HD* = HeaD.

CGN2.0 represents elided constituents by so-called SECONDARY EDGES between the remnant and the root node where the constituent is supposed to be elided. Secondary edges are represented by curved edges in tree diagrams provided by TIGERSearch. The grammatical function of the elided constituent labels the secondary edge.

TIGERSearch provides a specific search pattern for secondary edges (“>~”), possibly expanded by the grammatical function annotated at the edge. This allows accurate extraction of syntactic trees embodying various types of coordinate ellipsis. In CGN2.0, secondary edges are not directed (cf. TIGER). Nevertheless, the remnant can be determined by a search pattern in TIGERSearch. For both end nodes of a secondary edge, the grammatical function of the secondary edge is compared to the edge label of the node’s incoming edge. The matching one is the remnant. At the node the constituent is supposed to be elided, this function does not exist because it is supposed to be elided. One exception is the MODifier function that may occur repeatedly. These cases were inspected manually. The detailed proportions of different CCE types are reported in Section 3.1.3.

Appendix C: TIGER

In TIGER, edges are labeled with grammatical functions—similarly to CGN2.0. Edge labels are represented as squared boxed at the edges in tree diagrams provided by TIGERSearch (see, e.g., Figure 3 in Section 3.2.1). Nodes carry either lexical information or a phrasal category. In TIGER’s syntactic trees, different types of coordination are distinguished. The relevant ones for CCE are:

- CS: coordinated finite clauses,

- CVP: coordinated verb phrases (nonfinite clauses), and
- CVZ: coordinated infinitival clauses (VPs) with the verb preceded by *zu* ‘to’ (as in *zu tun* ‘to do’).

All those nodes dominate edges labeled *CJ* for conjunct and *CD* for coordinating conjunction. Such edges end in S, VP or VZ nodes. Abbreviations for edge labels in a clause are, e.g., SB = SuBject, HD = HeaD, OC = Object Complement.

Like in CGN2.0, TIGER represents elided constituents by so-called SECONDARY EDGES between the remnant and the root node where the constituent is supposed to be elided. Secondary edges are represented by curved edges in TIGERSearch tree diagrams. The grammatical function of the elided constituent labels the secondary edge.

In TIGER, secondary edges are DIRECTED from the remnant to the root node of which the constituent is supposed to be a child. There are FORWARD and BACKWARD secondary edges. All backward secondary edges point out BCR. Forward secondary edges require more fine-grained search patterns. For instance, the edge label HeaD at a forward secondary edge points to Gapping instances. The detailed proportions of different CCE types are reported in Section 3.2.3.

Appendix D: VERBMOBIL

In VERBMOBIL, three node types are differentiated. Nodes carry either lexical information or a phrasal category—similarly to CGN2.0 and TIGER. Moreover, topological field information is spelled out in so-called FIELD NODES:

- VF = VORFELD ‘Forefield’ with the special types listed in Table D-1 in more detail: MVC and MVCN which occur in clauses with a Complementizer (C), viz. as CMVC and CMVCN,
- MF = MITTELFELD ‘Midfield’,
- NF = NACHFELD ‘Endfield’,
- LK = LINKE SATZKLAMMER ‘left sentence bracket’, and
- VC = Verb Complex.

Coordination type	Composition pattern
MN	MF+NF
CM	C+MF
MVC/CMVC	MF+VC/C+MF+VC
MVCN/CMVCN	MF+VC+NF/C+MF+VC+NF
LKM	LK+MF
LKMN	LK+MF+NF
LKMVC	LK+MF+VC
LKMVCN	LK+MF+VC+NF
LKN	LK+MF+NF
VCN	VC+NF
VLKM	VF+LK+MF
VLKMVC	VF+LK+MF+CV

Table D-1. Node labels of conjuncts denoting clausal coordination in VERBMOBIL. A coordination type (in column 1) represents a node label. The composition pattern in column 2 spells out which subtree types (i.e. node labels at the root of the subtree) it dominates.

Edges are labeled in the VERBMOBIL corpus. Such labels are represented as squared boxed at the edges in TIGERSearch tree diagrams (see, e.g., Figure 4 in Section 3.2.2). All edges to field nodes are labeled with a dash. Within a clause, grammatical functions are represented as edge labels. For instance, ON = Subject and OA = Direct Object.

Clausal coordination is spelled out at the level of field nodes. A node labeled *FKOORD*, which represents complex field coordination, dominates nodes of the types listed in column one of Table D-1. The second column of Table D-1 describes the composition pattern, i.e. the nodes which have to occur as children of that node. The symbol ‘+’ separates the children’s types in the list.

Notice that, unlike ALPINO, CGN2.0 and TIGER, VERBMOBIL provides NO specification relating remnant and elided constituent. Accordingly, we first retrieved sentences with clausal coordination. Within this set, all sentences were manually classified. For instance, MN could be a case of FCR or SGF. The detailed proportions of different CCE types are reported in Section 3.2.3.

References

- Arnold, J.E., Wasow, T., Asudeh, A. & Alrenga, P. 2004. Avoiding Attachment Ambiguities: The Role of Constituent ordering. *Journal of Memory and Language*, 51:1, 55–70.
- Bell, A. 1984. Language Style as Audience Design. *Language in Society* 13(2): 145–204.
- Borsley, R.D. 2005. Against ConjP. *Lingua*, 115, 461–482.
- Boston, M.F., Hale, J.T., Vasishth, S. & Kliegl, R. In Press. Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*.
- Branigan, H.P., Pickering, M.J. & Cleland, A.A. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75, B13–B25.
- Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G. & Uszkoreit, H. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation*, 2, 597–620.
- Brennan, S.E. & Hanna, J.E. 2009. Partner-Specific Adaptation in Dialog. *Topics in Cognitive Science* 1, 274–291.
- Culicover, P.W. & Jackendoff, R. 2005. *Simpler Syntax*. Oxford: Oxford University Press.
- Ferreira, F. 2000. Syntax in Language Production: An Approach Using Tree-Adjoining Grammars. In Wheeldon, L. (ed.). *Aspects of Language Production*. Cambridge MA: MIT Press.
- Frank, R. & Badecker, W. 2001. Modeling syntactic encoding with incremental Tree Adjoining Grammar. In *Proceedings of the 14th Annual CUNY Conference on Human Sentence Processing*, Philadelphia PA.
- Greenbaum, S. & Nelson, G. 1999. Elliptical clauses in spoken and written English. In Collins, P. & Lee, D. (eds.). *The clause in English*. Amsterdam: Benjamins.
- Harbusch, K. & Kempen, G. 2007. Clausal coordinate ellipsis in German: The TIGER treebank as a source of evidence. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007)*, Tartu, Estonia.
- Harbusch, K. & Kempen, G. 2009a. A treebank study of clausal coordinate ellipsis in spoken and written language. In *Proceedings of the 15th Annual Conference on Architectures and Mechanisms of Language Processing (AMLaP2009)*, Barcelona, Spain.
- Harbusch, K. & Kempen, G. 2009b. Clausal Coordinate Ellipsis and its varieties in spoken and written German: A study with the TüBa-D/S Treebank of the VERBMOBIL corpus. In *Proceedings of the 8th International Workshop in Treebanks and Linguistic Theories (TLT8)*, Milano, Italy.
- Harbusch, K. & Kempen, G. 2009c. Generating clausal coordinate ellipsis multilingually: A uniform approach based on postediting. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, Athens, Greece.
- Hartmann, K. 2002. *Right Node Raising and Gapping: Interface Conditions on Prosodic Deletion*. Philadelphia/Amsterdam: John Benjamins.
- Haywood, S., Pickering, M.J. & Branigan, H.P. 2005. Do speakers avoid ambiguity in dialogue? *Psychological Science*, 16, 362–366.
- Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H. & Trushkina, J. 2004. Recent Developments in Linguistic Annotation of the TüBa-D/Z treebank. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT3)*, Tübingen, Germany.
- Höhle, T.N. 1983. Subjektlücken in Koordinationen. Unpublished manuscript, University of Cologne.
- Höhle, T.N. 1986. Der Begriff “Mittelfeld”, Anmerkungen über die Theorie der topologischen Felder. In Schöne, A. (ed), *Kontroversen, alte und neue: Akten des 7. Internationalen Germanisten-Kongresses*, Göttingen 1985, Band 3, 329–340. Tübingen: Niemeyer.

- Höhle, T.N. 1990. Assumption about asymmetric coordination in German. In Mascaró, J. & Nespó, M. (eds.). *Grammar in Progress: Glow Essays for Henk van Riemsdijk*, 221–235. Dordrecht: Foris.
- Hudson, R. 1976. Conjunction Reduction, Gapping, and Right-Node-Raising. *Language* 52:535–562.
- Kathol, A. 2001. Linearization vs. phrase structure in German coordination constructions. *Cognitive Linguistics*, 10:4, 303–342.
- Kempen, G. 2009. Clausal coordination and coordinate ellipsis in a model of the speaker. *Linguistics*, 47, 653–696.
- Kempen, G. & Hoenkamp, E. 1982. Incremental sentence generation: implications for the structure of a syntactic processor. In Horecky, J. (ed.). *Proceedings of the Ninth International Conference on Computational Linguistics (COLING)*, Prague. Amsterdam: North-Holland.
- Kempen, G. & Hoenkamp, E. 1987. An incremental procedural grammar for sentence formulation. *Cognitive Science: A Multidisciplinary Journal*, 11: 2, 201–258.
- König, E. & Lezius, W. 2003. *The TIGER language: A Description Language for Syntax Graphs, Formal Definition*. Tech. Rep. IMS, University of Stuttgart, Germany.
- Kübler, S., Hinrichs, E., Maier, W. & Klett, E. 2009. Parsing coordinations. *Proceedings of EACL 2009*, Athens, Greece.
- Lechner, W. 2004. *Ellipsis in Comparatives*. Berlin: Walter de Gruyter.
- Lemnitzer, L. & Zinsmeister, H. 2006. *Korpuslinguistik: Eine Einführung*. Tübingen: Narr Studienbücher.
- Levelt, W.J.M. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Meyer, C.F. 1995. Coordination Ellipsis in Spoken and Written American English. *Language Sciences*, 17, 241–169.
- Prince, A. & Smolensky, P. 1993/2004. *Optimality Theory: Constraint interaction in Generative Grammar*. Rutgers University and University of Colorado at Boulder: Technical Report RuCCSTR-2, available as ROA 537-0802. Revised version published by Blackwell, 2004.
- Purver, M. & Kempson, R. 2004. Incremental Context-Based Generation for Dialogue. In *Proceedings of the 3rd International Conference on Natural Language Generation (INLG04)*, Careys Manor, UK.
- Reich, I. 2007. Toward a Uniform Analysis of Short Answers and Gapping. In Schwabe, K. & Winkler, S. (eds.). *On Information Structure, Meaning and Form*. Amsterdam: John Benjamins. 467–484.
- Reich, I. 2008. From discourse to “odd coordinations”: On asymmetric coordination and subject gaps in German. In Fabricius-Hansen, C. & Ramm, W. (eds.). *‘Subordination’ versus ‘coordination’ in sentence and text: A cross-linguistic perspective*. Amsterdam: Benjamins.
- Sag, I.A., Wasow, T. & Bender, E.M. 2003. *Syntactic Theory: A formal introduction*. Stanford CA: CSLI publications [Second Edition.]
- Seginer, Y. 2007. Fast Unsupervised Incremental Parsing. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Prague, Czech Republic.
- Steedman, M. 1990. Gapping as Constituent Coordination. *Linguistics and Philosophy*, 13, 207–263.
- Steedman, M. 2000. *The syntactic process*. Cambridge MA: MIT Press.
- Stegmann, R., Telljohann, H. & Hinrichs, E. 2000. *Stylebook for the German Treebank in Verbmobil*. Saarbrücken: DFKI Rep. 239.
- te Velde, J.R. 2006. *Deriving Coordinate Symmetries*. Amsterdam: Benjamins.
- Uit den Boogaart, P.C. 1975. *Woordfrequenties in geschreven en gesproken Nederlands*. Werkgroep Frequentie-onderzoek van het Nederlands. Utrecht: Oosthoek, Scheltema & Holkema.
- van der Beek, L., Bouma, G., Malouf, R. & van Noord, G.-J. 2002. The Alpino Dependency Treebank. In *Computational Linguistics in the Netherlands (CLIN 2001)*. Amsterdam: Rodopi.

- van Eerten, L. 2007. Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, 12:3, 194–215.
- van Oirsow, R.R. 1987. *The syntax of coordination*. London: Croom Helm.
- Wahlster, W. (ed.). 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer.
- Wunderlich, D. 1988. Some problems of coordination in German. In Reyle, U. & Rohrer C. (eds.). *Natural Language Parsing and Linguistic Theories*, Dordrecht: Reidel.
- Zinsmeister, H. 2006. Treebank Data as Linguistic Evidence? Coordination in TüBa-D/Z. *Pre-Proceedings of the International Conference on Linguistic Evidence*, Tübingen.