Clausal Coordinate Ellipsis and its Varieties in Spoken German: A Study with the TüBa-D/S Treebank of the VERBMOBIL Corpus

Karin Harbusch
University of Koblenz-Landau
Computer Science Department
Koblenz, Germany
harbusch@uni-koblenz.de

Gerard Kempen
Max Planck Institute
for Psycholinguistics
Nijmegen, The Netherlands
gerard.kempen@mpi.nl

Abstract

Grammar rules for Clausal Coordinate Ellipsis (CCE) are based nearly exclusively on linguistic judgments (intuitions). For German, the extent to which grammar rules based on this type of empirical evidence generate all and only CCE structures that populate text corpora, has only been explored with the TIGER treebank of written newspaper text. How well these rules fit spoken German is unknown. In this paper, we study the applicability of judgment-based CCE rules to spontaneously spoken German by means of the TüBa-D/S treebank, which is based on dialogues for appointment scheduling and travel planning from the VERBMOBIL project. The judgment-based CCE rules are shown to hold nearly equally well for spoken as for written text: The proportion of deviations from the rules are virtually identical—less than 3% of the utterances/sentences that include a clausal coordination (compared to about 1% in the TIGER treebank). Moreover, the relative frequencies in VERBMOBIL of four main CCE types distinguished in the literature reveal a pattern that resembles the pattern observed in CGN2.0, the Corpus of Spoken Dutch.

1 Introduction

Coordinating conjunctions often license syntactic constituents to be elided from one conjunct if they have a (nearly) identical counterpart in another conjunct. Example (1), taken from the TüBa-D/S[1] treebank, exhibits "forward" elision of *viertel vor zwölf könnte* combined with "backward" elision of *abholen*. The presumed ellipsis sites are indicated by dots. At those sites,

[1] TüBa-D/S contains 38,228 sentences with about 380,000 word tokens (Stegmann et al. [16]) collected in the VERBMOBIL project (see Wahlster [21]). Here, a sentence refers to a complete dialogue turn by a speaker and thus often consists of several main clauses. The dialogue partners schedule appointments and set up traveling plans. In the following, we refer to the TüBa-D/S as the *"VERBMOBIL"* treebank in order to avoid confusion with the TüBa-D/Z treebank (Hinrichs et al. [7]) which is a corpus of German newspaper texts currently comprising about 22,000 sentences taken from the Wissenschafts-CD of "die tageszeitung"—henceforth called the "*TAZ*" treebank.

the elliptical conjuncts may be said to BORROW overtly mentioned counterparts from the parallel conjunct. The 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.)

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

Grammar rules for CLAUSAL COORDINATE ELLIPSIS (CCE) are based nearly exclusively on linguistic judgments (intuitions). In Harbusch & Kempen [3], we investigated how well a set of judgment-based rules aiming to generate all CCE varieties in German is obeyed in written language. For the TIGER treebank of newspaper texts (Brants et al. [1]), we reported 99% accuracy. How well spoken language fits these rules is not known. In the following, we present a qualitative and quantitative study of CCE in the VERBMOBIL (TüBa-D/S) treebank of spoken German dialogues. To anticipate one of the main results, the judgment-based CCE rules accord with more than 97% of the CCE tokens. However, the error types overlap only partly with CCE errors in written German.

The paper is organized as follows. In Section 2, we present an overview of the main types of CCE and spell out the rule set proposed by Kempen [9] and Harbusch & Kempen [5]. In Section 3, we briefly describe corpus studies on coordinate structures in German, Dutch and English reported in the literature. In Section 4, we present our study on CCE in the VERBMOBIL treebank. Moreover, we compare the German results to English and Dutch findings. Finally, in Section 5 we draw some conclusions and mention a desideratum for future work.

2 The Main CCE Types and How to Generate Them

In the linguistic literature on coordinate syntactic structures (for overviews, see van Oirsow [19]; Steedman [14]; Sag et al. [14]; te Velde [17]; and Kempen [9]), one often distinguishes four main types of coordinate ellipsis:[2]

- GAPPING, with three special 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).

[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 recasts of clausal coordinations 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 logical rather than syntactic mechanism.

Table 1. Clausal coordinate ellipsis (CCE) types as specified by Harbusch & Kempen [5]. Column 1 mentions the names of the CCE types and in brackets their abbreviations. Column 2 illustrates the CCE types in terms of English examples. The distinctions apply to German as well. Column 3 summarizes the elision conditions we apply in our study. Struck-out text represents borrowings.

CCE type	**Examples**	**Elision conditions**
Gapping (g)	(2) *Ulf* ***lives*** *in Leipzig and his children* ~~***live***~~$_g$ *in Ulm*	Lemma identity of Verb & contrast of remnants
LDG ((g)$^+$g)	(3) *My wife* ***wants to buy*** *a car and my son* ~~***wants***~~$_g$ ~~***[to buy]***~~$_{gg}$ *a motorcycle*	Gapping conditions in a *superclause*
Sub-gapping (sg)	(4) *The driver* ***was*** *killed and the passengers* ~~***were***~~$_{sg}$ *severely wounded*	Gapping conditions & VP remnant in second conjunct
Stripping (str)	(5) *Mare* ***lives in Narva*** *and her children* ~~***[live in Narva]***~~$_{str}$ ***too***	Gapping conditions & only one non-Verb remnant
FCR (f)	(6) ***Since a year, Kees*** *lives in Aam and* ~~***[since a year, Kees]***~~$_f$ *works in Edam* (7) *Tokyo is the city [****where*** *Ota lives and* ~~***where***~~$_f$ *Kusuke works]$_S$*	Wordform identity & left-peripherality (within clause boundaries) of major clausal constituents
BCR (b)	(8) *John wrote one* ~~***article***~~$_b$ *and Mary edited two* ***articles***. (9) *Anja arrived before three* ~~***[o'clock]***~~$_b$ *and Maria* ~~***arrived***~~$_g$ *after four* ***o'clock***	Lemma identity & right-peripherality, possibly disregarding major constituent boundaries
SGF (s)	(10) *Into the wood went* ***the hunter*** *and* ~~***[the hunter]***~~$_s$ *shot a hare*	Form-identical Subject & first conjunct starting with Verb/Modifier/Adjunct & FCR applied if licensed

As summarized in column 3 of Table 1, all forms of *Gapping* are characterized by elision of the posterior member of a pair of lemma-identical Verbs. The position of this Verb need not be peripheral but is often medial, as in (2) through (5), and (9).[3] Every non-elided constituent (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.[4] Stated differently, the members of such a pair are CONTRASTIVE—in (2): the Subjects *Ulf* vs. *his children*, and the locative Modifiers *in Leipzig* vs. *in Ulm*. Notice that although the two tokens of *my* in (3) occupy comparable positions in the two conjuncts, it is not possible to elide one of them because all CCE construc-

[3] In our definitions of CCE types, we restrict ourselves to coordinations encompassing two conjuncts, called *anterior* (first, left) and *posterior* (second, right), respectively.

[4] In the following, we distinguish three identity relationships between constituents in coordinated conjuncts: lemma identity, wordform identity and coreferentiality. For *lemma identity,* only the lexical entries ('syntactic words') 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 lemmata.

tions except BCR respect major constituent boundaries. On the other hand, *were* in (4) can be elided from the posterior conjunct although it has no word-form-identical (but only a lemma-identical) anterior counterpart.

In 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 do not include a subordinating Conjunction—with the possible exception of the topmost clause). In (3), *my son* belongs to the main clause but *a motorcycle* to the infinitival complement clause. In SUBGAPPING, the posterior conjunct includes a remnant in the form of a nonfinite complement clause (VP; *severely wounded* in (4)). In STRIPPING, the posterior conjunct is left with one non-Verb remnant, often supplemented by a sentential Adverb such as *too* or *not.*

In FCR, elision affects the posterior token of a pair of left-peripheral strings consisting of one or more wordform-identical major constituents. In (6), the posterior tokens of *since a year, Kees* and *where,* respectively, belong to such pairs and are eligible for FCR.

BCR is almost the mirror image of FCR as it deletes the anterior member of a pair of right-peripheral lemma-identical word strings (*o'clock* in (9)); however, BCR may elide part of a major constituent—e.g. only the part *article* of the Direct Object in (8) and *o'clock* of the temporal Modifier *before three o'clock* in (9). In addition, it requires only lemma identity (cf. example (8)).

SGF can elide the Subject of the posterior conjunct—always 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 (11) where a Complement clause opens the anterior conjunct. (In well-formed SGF case (10), the initial constituent is an Adjunct.)[5]

(11) **Das Examen bestehen will er und* ~~*er*~~$_s$ *kann auch*
The exam to-pass wants he and can too
'He wants to pass the exam and will be able to as well'

3 Previous Corpus Work on CCE in German, Dutch and English

In a recent paper (Harbusch & Kempen [3]), we analyzed the incidence of clausal coordination and coordinate ellipsis in the TIGER treebank. TIGER contains 50,474 syntactically annotated sentences originating from a German newspaper corpus. Almost 43% of them (21,506 sentences include a coordinate structure of any type, and 7,194 sentences (33% of the latter) contain at least one clausal coordination. In total, 4,020 TIGER sentences contain at

[5]We also subsume under the heading of SGF cases like (i), where the anterior conjunct is a conditional subordinate clause. See Höhle [8] and Reich [13] for discussion of the affinity between this structure and SGF as defined here.

(i) *ja, dann reicht es ja, wenn wir ungefähr um neun losfahren würden und* ~~*wir*~~ *würden dann mittags dort ankommen*
'OK, then it suffices if we would leave at nine and would arrive there in the afternoon'

least one CCE token, distributed over the four main CCE types as follows: 2545 cases of FCR (63%), 678 Gapping tokens (17%), 384 SGF cases (10%), and 413 BCR tokens (10%).[6] Of these, 99% percent obey the rules of Table 1. Only some 40 sentences violate a borrowing rule but were judged at least marginally acceptable. These sentences embody four borrowing (elision) patterns that may be characterized as 'fringe deviations' from the intuition-based coordinate ellipsis rules: *overreduction, peripherality violations by little words, peripherality violations by content words or word groups*, and *sloppy gapping*.

For Dutch, we conducted a comparative study of CCE in written and spoken language ([4][6]). We explored the ALPINO treebank (van der Beek et al. [18]) consisting of 7,153 manually annotated syntactic structures from a newspaper corpus, and CGN2.0 (van Eerten [19]) with about 130,000 spoken sentences or dialogue turns from more than ten different domains. In written Dutch, the percentage of elliptical versions within the set of all clausal coordinations is three times higher than in spoken Dutch: 34% versus 11% (Harbusch & Kempen [4]). In each of the treebanks, Gapping and FCR together covers 92% of the CCE cases (with the remaining 8% more or less evenly distributed among SGF and BCR). However, the distribution of FCR and Gapping in the two Dutch treebanks differs widely. Whereas in written clausal coordinations Gapping accounts for only 10% of the CCE cases (with a large majority of 82% embodying FCR), in spoken clausal coordinations the incidence of Gapping is much higher: 31% (leaving 61% for FCR). These numbers are comparable to those observed in the German written and spoken corpora (the latter are reported in the next Section).

In two corpus studies into the incidence of CCE in spoken and written English, Meyer [12] and Greenbaum & Nelson [2] found that in written clausal coordinations, the proportion of elliptical versions is about twice as high as in spoken coordinations.

These findings suggest that there may be substantial cross-linguistic similarities, at least as far as the Germanic languages are concerned, with respect to the frequencies of CCE and CCE types in spoken texts and in written texts.

4 Clausal Coordinate Ellipsis in the VERBMOBIL Treebank

After an outline of the methodology of the corpus study, we report on the accuracy of the elision rule set and the error classes found in the VERBMOBIL corpus. Finally, we compare our frequency results to the findings for Dutch and English reported in the previous Section.

[6] In a quantitative study into the German TAZ treebank, Zinsmeister [22] found 8,133 sentences (37% of the total number of sentences) that include a coordination of syntactic constituents of any type (marked by the edge label KONJ). She only reports one number dealing with CCE types: 83 sentences with SGF—i.e about 1% of all coordinations. This percentage is comparable to the proportion of SGF cases in TIGER: the 384 cases we observed there, make up less than 2% of the total number of coordinations.

4.1 Methodological Issues

The VERBMOBIL treebank is encoded in the same manner as the TAZ corpus (Hinrichs et al. [7]) but rather differently from TIGER (see Lemnitzer & Zinsmeister [11], page 82, for a comparison of the tag sets).

In TIGER, coordinate elisions are explicitly marked by SECONDARY EDGES, i.e. edges that run from the root node of a remnant—a borrowed string—to the root node of the structure that borrows the remnant as a child node. The edge label indicates the grammatical function that the borrowed remnant fulfils in the borrowing structure. These encodings enable the composition of search queries that automatically retrieve CCE structures (by means of TIGERSearch; König & Lezius [10]). Moreover, they support semi-automatic classification of CCE types and verification of the elision rules.

Secondary edges do not occur in VERBMOBIL trees. Therefore, we manually inspected all clausal coordinations for CCE, classified them for CCE type, and marked all rule violations. Additionally, we accessed and checked all dialogue turns from which the CCE tokens originate and ruled out any false alarms generated by the search queries. This procedure enabled us to compare the frequency data for written text in TIGER with the frequencies of spoken text in VERBMOBIL.

The VERBMOBIL frequency counts proceeded in two steps. First, we collected all clausal coordinations consisting of one or more incomplete conjuncts—incomplete in the sense that some constituent(s) seemed to be missing. Utterances that we judged to be ill-formed due to a self-correction by the speaker, like sentence (12), were left out of consideration. This also happened to over 100 cases which include a left-dislocated constituent followed by a resumptive pro-form—*das* 'that' in (13)—, which one could tentatively analyze as opening the posterior conjunct of an asyndetic coordination, with deletion of a right-peripheral string in the first conjunct, as in BCR.

(12) *da hat da hätte ich auch Zeit*
then have then would-have I too time
'then I would have time as well'

(13) *aber siebzehnter, achtzehnter* ~~*ginge*~~ *das ginge* .
'but seventeenth, eighteenth that would-be-possible'

We also ruled out all cases, which we judged to result from plausible *conceptual inference* rather than from borrowing licensed by a coordinating conjunction. In (14), the Adverb 'then' probably modifies not only the anterior but also the posterior conjunct. However, the absence of does not render the second conjunct incomplete. Hence, we classified (14) as a well-formed case of FCR with borrowing of the Personal Pronoun 'we' only. (For details regarding conceptual inference, see Harbusch & Kempen [3].)

(14) *wir fliegen dann am elften und* ~~*wir*~~ *bleiben für zwei Tage*
,'we fly then on-the eleventh and stay for two days'

Importantly, we only considered structures that do not allow an alternative analysis as a nonclausal coordination of NPs, PPs, APs, etc. For instance, sentences (15) and (16) were discarded due to the possibility of analyzing

them as PP-coordination (instead of as a combination of BCR and FCR—cf. example (1)). Importantly, however, just as in our TIGER study, we included nonclausal coordinations into the CCE counts if the posterior conjunct follows the clause-final Particle or Verb of the anterior conjunct—see (17) and (18) for an illustration. In the classification of CCE types, we group them together with the Stripping variant of Gapping.

To prevent a misunderstanding, whenever a VERBMOBIL utterance of the type discussed here had been encoded explicitly either as a discontinuous structure or as Stripping/Gapping, we adopted this choice. (In (17), PP *nach Hannover* was encoded as an extraposed part of the NP headed by *Reise*; and *und zwar* was encoded as a discourse marker rather than as a syntactic node.) However, very often the encodings left the choice between discontinuous structure vs. Stripping/Gapping open.

(15) *dann sage ich meiner Sekretärin [wegen der Bahnkarten und wegen dem Hotel]$_{PP}$ Bescheid*
'Then I'll inform my secretary about the tickets and about the hotel'

(16) *ich war schon ein paar Mal [in Hannover und zwar in dem Hotel Loccumer-Hof]$_{PP}$*
'I was already a few times in Hannover, namely in hotel Loccumer-Hof'

(17) *ich habe eine Reise vor, und zwar nach Hannover*
I have a trip in-mind namely to Hannover

(18) *schauen wir noch, ob wir noch ins Theater gehen oder in ein Kino*
look we also whether we also to-the theater go or to a cinema
'let's also look whether we go to the theater or to a cinema'

In the second step, we classified the CCE tokens according to CCE type. Like in our TIGER study, when a sentence embodies several CCE constructions, we counted each of them separately. Recall that, in VERBMOBIL, sentence numbers were assigned to entire dialogue turns, which often include several (main) clauses. Sentence (19), for example, features two FCR cases, actually borrowing different left-peripheral strings.

(19) *wenn ich da nicht da wäre und ~~wenn~~$_f$ er in meinem Büro sitzen würde und ~~wenn er~~$_f$ Däumchen drehen würde*
'if I wouldn't be in and he would sit in my office and kick his heels'

When a CCE instance could be ranged under more than one type, we followed the encodings the TIGER treebank as much as possible. For cases like (20), for instance, we chose the FCR analysis although the sentences can be viewed as nonelliptical coordinations of infinitival clauses.

(20) *Oder möchten Sie sparen und ~~möchten~~ ~~Sie~~ das Doppelzimmer nehmen?*
'Or would you like to save money and take a double room?'

Finally, while carrying out these steps, we sometimes needed to 'clean up' the sentence materials, for instance, to remove interjections or to insert words that were missing for reasons clearly unrelated to coordinate ellipsis. In (21), the Subject NP *Sie* 'you' seems to be missing after the second token of *wenn* 'if'. (Given this reconstruction, the sentence is analyzed as a Stripping variant of Gapping.) In (22), the speaker interrupts the PP headed by *zwischen* 'be-

tween', inserts a series of editing terms and interjections (the string between vertical bars), and resumes with *vielleicht* 'perhaps' and a revised PP. We disregarded the string between bars and interpreted the sentence as Gapping, with a contrast between *sehr gut* 'very well' and *vielleicht* on the one hand, and between the original and the revised dates on the other.

(21) *wenn Sie möchten, wenn (Sie) sich vielleicht das Museum-für-Hamburgische-Geschichte ansehen oder ~~wenn~~ ~~(Sie)~~ ~~sich~~ die Kunsthalle, die neu eröffnet worden ist ~~ansehen~~*
'If you like, if (you) visit the historical museum of Hamburg or the art gallery which has just reopened'

(22) *sehr gut passen würde es mir zwischen siebten Mai und || nee, ach doch das ist doch nicht gut das ist gar || vielleicht ~~passen~~ ~~würde~~ ~~es~~ ~~mir~~ zwischen dem achten Juni und elften Juni*
'very well would suit me between the 7th of May and || no, oh yeah, this is not good, this is even || perhaps between the 8th and 10th of June'

4.2 CCE Error Types in the VERBMOBIL Treebank

We found 3,713 VERBMOBIL sentences (or rather dialogue turns) with at least one clausal coordination (including asyndetic ones). This set includes 1,314 sentences (35%) with at least one CCE token. (See the next Subsection for the frequencies of the individual CCE types.) In the remainder of this Section, we concentrate on two questions: How well does spoken language obey the elision rules of Section 2, compared to written language? And to what extent do the error types observed in spoken language overlap with those seen in written language?

In VERBMOBIL, we identified 35 CCE tokens that violated a judgment-based CCE rule. That is, less than 3% of the sentences was ill-formed—a proportion not substantially different form the 1% deviations we reported earlier for written German. We find this somewhat surprising, given that spoken language is supposed to be more error-prone than written language: The speaker is under higher time pressure, has no external memory, and cannot easily hide away editing actions from the audience. On the other hand, the conceptual and grammatical structure of spoken sentences tends to be much simpler than that of written sentences (cf. average sentence length).

Of the 35 CCE error tokens, only a minority could be ranged unequivocally under the error classes we distinguished in Harbusch & Kempen [3]. We classified 13 tokens as OVERREDUCTIONS. In errors of this type, the elision process cuts into a major clause constituent functioning as remnant, with the consequence that only part of the remnant survives. In Stripping/Gapping example (23), the speaker failed to repeat *Fahrschein* 'ticket' in the posterior conjunct (in addition to deleting the Prepositon *aus* 'leaving' at the end of the PP). In (24), also classified as Stripping/Gapping, the Preposition+Article *ins* was left out in front of *Schauspiel* 'theater play'.

(23) *... ob es möglich ist einen Fahrschein von Dammtor aus zu*
whether it possible is a ticket from Dammtor leaving to
bekommen und nicht vom Hauptbahnhof
get and not from main-station
'... whether it is possible to get a ticket leaving from Dammtor and not one leaving from main station'

(24) *vielleicht könnten wir ins Theater gehen, Schauspiel, oder in eine Oper*
maybe can we to-the theater go play or to an opera
'maybe we can go to the theater, the playhouse or the opera'

Another error class we had dubbed SLOPPY GAPPING: The verb elided from the posterior conjunct is used with a subcategorization frame different from that of its counterpart in the anterior conjunct. For instance, the Verb *werden* 'be' is used as passive Auxiliary in one clausal conjunct and as a copula Verb in another. In VERBMOBIL, we found 4 coordinations that arguably feature this type of error. In (25), *mögen* 'like' functions first as modal Verb, then as transitive Verb. In the anterior conjunct of (26), *mögen* is used as intransitive Verb; the posterior conjunct requires the modal Verb *mögen.*

(25) *weiß nicht, in die Oper möchte ich grade nicht gehen, oder ins Konzert, aber vielleicht irgendwas kleines gemütliches Treffen*
'don't know, I wouldn't like to go to the opera or to a concert, but perhaps something small, (a) cosy meeting'

(26) *und ich möchte dann schon am fünfzehnten noch mal schnell ins Büro und schauen was sich da so angesammelt hat*
'and then already on the 15th I'd like (to go) quickly into the office and look what has been piling up there'

In both overreduction and sloppy gapping, the speaker presumably does not accurately take into account the constraints imposed by the syntactic shape of the anterior conjunct.

Of the two remaining error classes (peripherality violations by little words, or by content words or word groups), we could not find a single unequivocal instance in VERBMOBIL. However, in 18 dialogue turns we spotted a mixed bag of other imperfections. In (27), for instance, the Particle *zurück* 'back' does not have a contrastive counterpart; prefixing *fahren* with Particle *hin* 'away' would have made for a perfect Gapping structure. In (28), an FCR case gone awry, the right conjunct needs an initial adverbial modifier like *da* 'there' but the left conjunct only has the Subject NP *das* 'that' on offer.

(27) *ich würde also gern am am Montag vormittags fahren und*
I would thus gladly on-the on-the Monday morning travel and
am Freitag nachmittags zurück ...
on-the Friday afternoon back
'I'd like to travel Monday morning and (come) back Friday afternoon'

(28) *das ist direkt am Hauptbahnhof und kostet das Einzelzimmer*
that is directly at-the main-station and costs the single-room
einhundert und neunundzwanzig Mark.
one-hundred and twenty-nine Mark
'that is directly at the main station and (there) a single room is 129 DM'

Quite a few dialogue turns consist mainly of utterances in telegram style. We classified some 25 exemplars as clausal coordinations if at least one of the conjuncts includes a finite verb. In the majority of those coordinations (some asyndetic), the clause-initial topic position of both conjuncts is empty: 'topic drop'; cf. (29) and (30). As the fillers of both topic positions are identical, we provisionally assume that the anterior filler is reconstructed from context, and that the posterior filler is borrowed from the—then reconstructed—anterior filler. If correct, this entails that we need not classify these cases as CCE errors but as legal CCEs.

(29) *startet Bonn Hauptbahnhof um acht Uhr fünfundvierzig und kommt an*
leaves Bonn main-station at 8 hour 45 and arrives
in Hannover Hauptbahnhof um zwölf Uhr vier
in Hannover main-station at 12 hour 4

(30) *liegt zentral, hat Hallenbad und Fitneßraum*
is located centrally has indoor-pool and fitness-room

4.3 CCE Frequencies in the VERBMOBIL Treebank, and Comparison with Other Studies

At the end of Section 3, we hypothesized that the CCE frequencies in spoken and written German, Dutch and English texts would exhibit similar patterns. The evidence obtained from the VERBMOBIL corpus supports this hypothesis. In all three languages, the proportion of CCE sentences within the total set of coordinated clauses is substantially higher in the spoken than in the written modality.

Table 2 shows the relative frequencies in German and Dutch of the four CCE types we distinguish. It reveals striking within-modality and within-language similarities. In the spoken modality, the incidence of Gapping is higher than in written language, mainly at the expense of FCR. In the German treebanks, BCR and SGF are well represented (in particular SGF) whereas in the Dutch corpora they live a somewhat marginal existence.

Table 2. Relative frequencies of the four types of CCE, expressed as percentages of the total set of sentences exhibiting CCE.

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

5 Discussion

We investigated to which extent Grammar rules for Clausal Coordinate Ellipsis, which are nearly exclusively based on linguistic judgments (intuitions), hold for spoken German. We followed the lead of a similar study conducted recently with the TIGER treebank for written German, using the TüBa-D/S

treebank, which is based on dialogues for appointment scheduling and travel planning from the VERBMOBIL project. After having presented a set of judgment-based CCE rules for four main CCE types distinguished in the literature, we showed that these rules fit spoken text nearly equally accurately as written text: The proportions of utterances deviating from the rules are very similar—less than 3% of the spoken sentences that include a clausal coordination, compared to about 1% of the written sentences in the TIGER treebank. However, the rule violations in the spoken corpus turned out to be of a rather different nature than those in the written corpus. Furthermore, we found that the relative frequencies in VERBMOBIL of the four main CCE types reveal a pattern that strongly resembles the patterns observed in the CGN2.0 treebank, the Corpus of Spoken Dutch.

We conclude not only that parsers and generators for spoken German can rely on the intuition-based rule systems for CCE, in particular rules such as described in Section 2 above, but also that their performance can profit from measures that allow for the fringe deviations observed in Section 4.

In future work we hope to provide a psycholinguistic explanation for the frequency/error patterns obtained in the present study and its predecessors.

References

[1] Brants, S., Dipper, S., Eisenberg, P., Hansen-Schirra, S., König, E., Lezius, W., Rohrer, C., Smith, G. and Uszkoreit, H. (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation, 2,* 597-620.

[2] Greenbaum, S. and Nelson, G. (1999). Elliptical clauses in spoken and written English. In: Collins, P. and Lee, D. (Eds.). *The clause in English.* Amsterdam: Benjamins.

[3] Harbusch, K. and 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.

[4] Harbusch, K. and 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.

[5] Harbusch, K. and Kempen, G. (2009b). 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.

[6] Harbusch, K. and Kempen, G. (2009c). Incremental sentence production inhibits clausal coordinate ellipsis: A comparison of spoken and written language. In: *Proceedings of the Workshop on Incrementality in Verbal Interaction*, Bielefeld.

[7] Hinrichs, E., Kübler, S., Naumann, K., Telljohann, H. and 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* (*TLT04),* Tübingen.

[8] Höhle, T.N. (1990). Assumption about asymmetric coordination in German. In: Mascaró, J. and Nespor, M. (Eds.), *Grammar in Progress: Glow Essays for Henk van Riemsdijk*, 221–235. Dordrecht: Foris.

[9] Kempen, G. (2009). Clausal coordination and coordinate ellipsis in a model of the speaker. *Linguistics, 47,* 653-696.

[10] König, E. and Lezius, W. (2003). *The TIGER language: A Description Language for Syntax Graphs, Formal Definition*. Tech. Rep. IMS, University of Stuttgart.

[11] Lemnitzer, L. and Zinsmeister, H. (2006). *Korpuslinguistik: Eine Einführung.* Tübingen: Narr Studienbücher.

[12] Meyer, C.F. (1995). Coordination Ellipsis in Spoken and Written American English. *Language Sciences, 17*, 241-169.

[13] Reich, I. (2008). From discourse to ‘‘odd coordinations’’: On asymmetric coordination and subject gaps in German. In: Fabricius-Hansen, C. and Ramm, W. (Eds.), *‘Subordination’ versus ‘coordination’ in sentence and text: A cross-linguistic perspective.* Amsterdam: Benjamins.

[14] Sag, I.A., Wasow, T. and Bender, E.M. (2003). *Syntactic Theory: A formal introduction*. Stanford CA: CSLI publications [Second Edition.]

[15] Steedman, M. (2000). *The syntactic process.* Cambridge MA: MIT Press.

[16] Stegmann, R., Telljohann, H. and Hinrichs, E. (2000). S*tylebook for the German Treebank in Verbmobil*. Saarbrücken: DFKI Rep. 239.

[17] te Velde, J.R. (2006). *Deriving Coordinate Symmetries*. Amsterdam: Benjamins.

[18] van der Beek, L., Bouma, G., Malouf, R. and van Noord, G.-J. (2002). The Alpino Dependency Treebank. In: *Computational Linguistics in the Netherlands CLIN 2001*. Amsterdam: Rodopi.

[19] van Eerten, L. (2007). Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, *12*, *3*, 194-215.

[20] van Oirsow, R.R. (1987). *The syntax of coordination*. London: Croom Helm.

[21] Wahlster, W. (Ed.) (2000). *Verbmobil: Foundations of Speech-to-Speech Translation.* Berlin: Springer.

[22] Zinsmeister, H. (2006). Treebank Data as Linguistic Evidence? Coordination in TüBa-D/Z. *Pre-Proceedings of the International Conference on Linguistic Evidence,* Tübingen.