Clausal coordinate ellipsis in German: The TIGER treebank as a source of evidence

Karin Harbusch
University of Koblenz-Landau
Computer Science Department
Universitätsstraße 1
56070 Koblenz, GERMANY
harbusch@uni-koblenz.de

Gerard Kempen
Max Planck Institute for Psycholinguistics
PO Box 310
6500 AH Nijmegen
THE NETHERLANDS
gerard.kempen@mpi.nl

Abstract

Syntactic parsers and generators need high-quality grammars of coordination and coordinate ellipsis—structures that occur very frequently but are much less well understood theoretically than many other domains of grammar. Modern grammars of coordinate ellipsis are based nearly exclusively on linguistic judgments (intuitions). The extent to which grammar rules based on this type of empirical evidence generate all and only the structures in text corpora, is unknown. As part of a project on the development of a grammar and a generator for coordinate ellipsis in German, we undertook an extensive exploration of the TIGER treebank—a syntactically annotated corpus of about 50,000 newspaper sentences. We report (1) frequency data for the various patterns of coordinate ellipsis, and (2) several rarely (but regularly) occurring 'fringe deviations' from the intuition-based rules for several ellipsis types. This information can help improve parser and generator performance.

1 Introduction

Coordinate structures often license elision of all but one of a set of syntactic constituents that express the same conceptual structure. In example (1) (next page), the conceptual structure underlying *my sister* belongs to the meaning of both conjuncts but is expressed overtly only in the anterior conjunct. The presumed ellipsis site is indicated by dots. At that site, the elliptical conjunct 'BORROWS' its overt counterpart from the parallel conjunct.

In this paper, we present frequency data for the various types of elliptical constructions in German—data extracted from the TIGER treebank (Brants *et al*., 2004). The frequencies can help improve generator and parser performance by guiding the selection of elision sites (in generation) and the reconstruction of elided materials (in parsing).

In the course of this project, we observed rare but nevertheless systematic deviations from ellipsis rules reported in the literature. These observations necessitate amendments to these rules.

In Section 2, we present an overview of the main phenomena of coordinate ellipsis. Section 3 characterizes the TIGER treebank. In Section 4, we report the key results from our treebank exploration and discuss implications for the grammar and for sentence parsing and generating. Finally, Section 5 outlines options for future work.

2 Coordinate ellipsis: the main phenomena

In the linguistic literature on coordinate syntactic structures (for overviews, see Van Oirsow, 1987; Johannessen, 1998; Steedman, 2000; Sag, Wasow & Bender, 2003; Te Velde, 2006; and Kempen, in press), one often distinguishes four main types of coordinate ellipsis:[1]

[1]We will not deal 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 deal with 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 a syntactic mechanism.

- Forward Conjunction Reduction (FCR),
- GAPPING, with three special variants called Long Distance Gapping (LDG), SUBGAPPING, and STRIPPING,
- Backward Conjunction Reduction (BCR; also known as Right Node Raising or RNR), and
- Subject Gap in clauses with Finite/Fronted verbs (SGF).

They are illustrated in the English sentences (1) through (7). The distinctions also hold for German.

(1) FCR*: My sister lives in Utrecht and ... works in Amsterdam*
(2) GAPPING*: Last year, John had an office in Leiden and ... Peter ... in Nijmegen*
(3) LDG*: My wife wants to buy a car and my son ... a motorcycle*
(4) SUBGAPPING: *The driver was killed and the passenger ... severely wounded*
(5) STRIPPING*: My sister lives in Utrecht and my brother ..., too*
(6) BCR*: Anne arrived before three ... and Susi left after four o'clock yesterday*
(7) SGF*: Why did you leave but didn't ... tell me?*

The main defining characteristics of these ellipsis types are as follows. Notice, in particular, the different borrowing patterns (described and empirically justified in detail by Kempen, in press).

- In FCR, the anterior and the posterior conjoined clauses each include an overt head verb (*lives* and *works* in (1)). Borrowing by the posterior conjunct is restricted to left-peripheral major constituents[2] shared by the conjuncts.
- In GAPPING, the posterior conjunct consists of one or more major constituents, each expressing a contrast with a major constituent in the anterior conjunct. The constituents of the posterior conjunct are often called REMNANTS. The posterior conjunct borrows obligatorily all and only those major constituents of the anterior conjunct that are non-contrastive, and this set must include the head verb (in (2): *last year*, *had* and *an office*). This characterization is also valid for LDG, Subgapping and Stripping (see below). An important exception applies to negation elements, which are not always borrowed and are usually repeated in the posterior conjunct:

  (8) *Hans wohnt nicht in Paris und Peter nicht in Rom*
  Hans lives not in Paris and Peter not in Rome
  'Hans doesn't live in Paris and Peter doesn't in Rome'.

- In LDG, the posterior conjunct consists of constituents whose left-hand counterparts belong to different clauses. *My son* in (3) is the counterpart of *my wife* in the main clause whereas *a motorcycle* pairs up with *a car* in the infinitival complement clause.
- SUBGAPPING is a special case of simple Gapping: the posterior conjunct includes one major constituent in the form of a non-finite complement clause ("VP"; *severely wounded* in (4)).
- STRIPPING is Gapping with the posterior conjunct consisting of one constituent only. This remnant is not a verb, and it is often supplemented by a modifier (such *too* in (5), *in particular*, or Ger. *zwar* 'more precisely').
- In BCR, the anterior conjunct borrows one or more—complete or partial—right-peripheral constituents from the posterior one (*o'clock* and *yesterday* in (6)).
- SGF is a coordination of MAIN clauses where the anterior conjunct exhibits subject-verb inversion (*did you* instead of *you did* in (7)), and the posterior conjunct borrows the anterior clause's subject NP. The posterior clause starts with the finite head verb, optionally borrowing the clause-initial (left-peripheral) modifier (if any—an adverbial phrase or clause, or a prepositional phrase). No other constituents are borrowable.

Modern grammars of coordinate ellipsis are based nearly exclusively on linguistic judgments (intuitions). The extent to which grammar rules based on this type of empirical evidence generate all and only the structures that populate text corpora, is unknown. The recent availability of the TIGER treebank (Brants *et al.*, 2004) enabled us to explore this question as part of a project on the development of a grammar and a generator for coordinate ellipsis in German and Dutch (Kempen, in press; Harbusch & Kempen, 2006).

[2] We use the term "major constituent" of a clause in a broad sense that includes head verb (main, copula or auxiliary), arguments (e.g. subject, direct and indirect object, and non-finite complement clause), adjuncts (adverbial modifier, including adverbial clause), and subordinating conjunctions (i.e. the complementizer in complement clauses—*that*, *whether*—or the subordinator in adverbial clauses—*while, although, when,* etc.

3 A corpus study of clausal coordinate ellipsis in German

3.1 TIGER: Characterization and annotation

The TIGER Treebank (Release 2) contains 50.474 German syntactically annotated sentences from a German newspaper corpus. As illustrated in Figures 1 and 2, TIGER's annotation scheme uses many clause-level grammatical functions (subject, direct and indirect object, complement, modifier, etc.; depicted as edge labels in the sentence diagrams). Important for present purposes, elided (i.e. borrowed) constituents in coordinate clauses are represented by so-called SECONDARY EDGES, also labelled with a grammatical function. This feature facilitates well-targeted automatic recognition and extraction of syntactic trees that embody various types of coordinate ellipsis. Secondary edges are represented by curved arrows in TIGER tree diagrams such as Figures 1 and 2.

In TIGER's syntactic trees, the following types of coordination are distinguished:

- CAC: coordinated adpositions,
- CAP: coordinated adjectival phrases,
- CAVP: coordinated adverbial phrases,
- CCP: coordinated complementizer phrases (subordinating conjunctions),
- CNP: coordinated noun phrases,
- CO: coordination of "unlikes", i.e. of different categories (e.g. an AP and a PP)
- CS: coordinated finite clauses,
- CVP: coordinated verb phrases (non-finite clauses), and
- CVZ: coordinated infinitival clauses (VPs) with the head verb preceded by *zu* 'to' (as in *zu tun* 'to do').

Within a coordinate structure, the conjuncts are dominated by a CJ edge, and the coordinating conjunction by a CD edge. In the current project, we focus attention on the three latter types: coordinated finite and non-finite (including infinitival) clauses.

The three bottom rows of Table 1 show that 7194 corpus sentences—about 14 percent—include at least one clausal coordination, and that in more than half of these (4046) one or more constituents have been elided and need to be borrowed from the other conjunct. According to Brants *et al.* (2004: p. 599), "secondary edges are only employed for the annotation of coordinated sentences and verb phrases"—CS, CVP and CVZ.[3] Nevertheless, secondary edges occasionally turn up as parts of non-clausal coordination types—see the shaded cells of Table 1. However, ellipsis in non-clausal coordinate structures is not annotated systematically.

We deployed the TIGERSearch tool (König & Lezius, 2003)

- to design queries that retrieve all clausal coordinations (whether elliptical or not), and
- to classify the elliptical ones (those including one or more secondary edges) into one of the seven (sub)types of clausal coordinate ellipsis.

We took into consideration all clausal coordinations, including asyndetic ones (lacking an overt coordinating conjunction), and those consisting of more than two conjuncts. To simplify the computational corpus explorations, we assume that the treebank does not contain sentences from which secondary edges are missing.

Table 1. Number of TIGER sentences that include one or more coordinations of the type mentioned in the first column. The two rightmost columns indicate how many sentences contain at least one secondary edge.

Coordination type	Total	With secondary edge	
		Forward	Backward
CAC	30	0	0
CAP	2170	2	1
CAVP	204	0	0
CCP	2	0	0
CNP	10282	0	3
CO	374	3	0
CPP	1250	5	2
CS	5607	3150	343
CVP	1564	466	86
CVZ	23	1	0

(9) *Monopole sollen geknackt und Märkte*
Monopolies should shattered and markets
getrennt werden
split be
'Monopolies should be shattered and markets split'

Figure 1 shows the tree diagram for example (9)—a clausal coordination which combines Subgapping

[3] We brought the TIGER structures annotated as CS, CVP and CVZ together under the heading of (non-)finite coordinated clauses. The left- and right-peripherality patterns of CVP and CVZ coordinations were checked by hand.

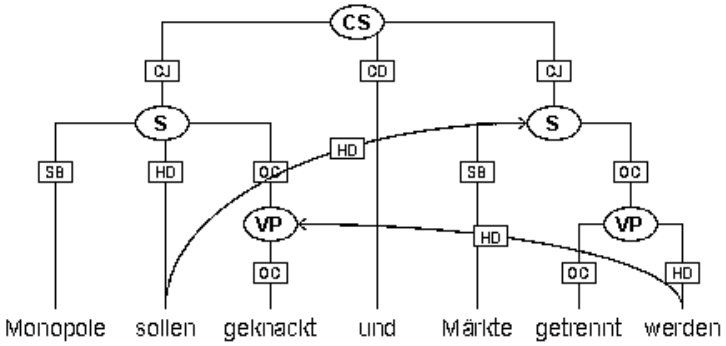

Figure 1. Tree diagram for example (9): Subgapping combined with BCR. The two remnants of the posterior clause are the NP *Märkte* and the VP (non-finite clause) *getrennt werden*. Abbreviations for edge labels: SB=subject, HD=head, OC=object complement, CD=coordinating conjunction, CJ=conjunct.

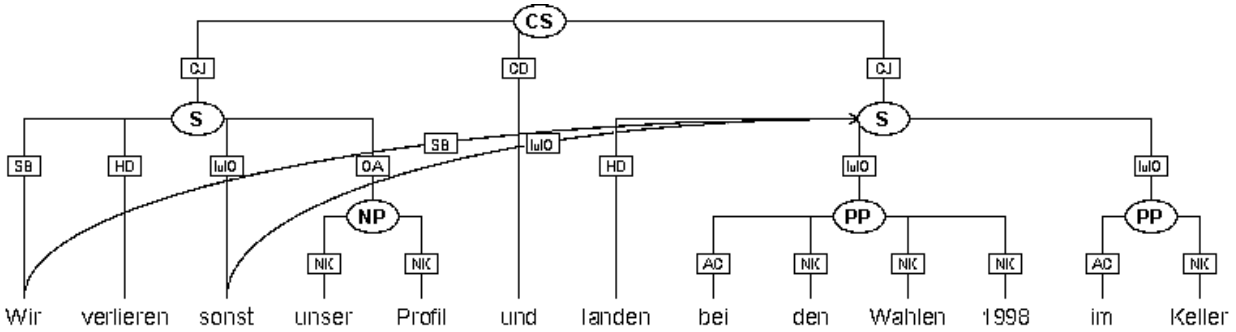

Figure 2. Tree diagram for FCR example (10): The posterior clause is headed by the overt finite verb *landen* and borrows its subject *wir* from the left. For the secondary edge dominating the adverbial modifier *sonst*, see the discussion in Section 3.2. Abbreviations: OA=direct object, NK=noun kernel/modifier, AC=adpositional case marker.

Table 2. Number of TIGER sentences with at least one clausal coordination, each sentence containing one or more secondary edges labelled with one of seven important grammatical functions. The total number of sentences with at least one clausal coordination (elliptical or non-elliptical) is shown within parentheses. Hence, the first number in a cell denotes a set of sentences that is a subset of the set denoted by the number in parentheses. The grey cells indicate borrowings that are either ruled out by the definition of the ellipsis type, or are entailed by the definition. E.g., SGF entails a secondary edge dominating the subject of the anterior clause, and rules out borrowings of constituents other than adverbial modifiers. The set of seven grammatical functions is not exhaustive because TIGER's annotation scheme distinguishes more grammatical functions than the seven listed here. As many TIGER sentences embody more than one clausal coordination, the numbers in a column do not add up to the total in the top row.

	Type of clausal coordinate ellipsis			
Borrowed (elided) constituent	**FCR N=2545**	**Gapping N=678**	**SGF N=384**	**BCR N=413**
Head verb of clause		678 (678)		22 (392)
Subject	1772 (2147)	208 (595)	384 (384)	27 (228)
Direct Object	10 (154)	6 (26)		1 (19)
Indirect Object	207 (1379)	55 (195)		24 (122)
Modifier	625 (1897)	197 (551)	157 (359)	73 (295)
Complementizer	433 (456)	9 (11)		0 (6)
Particle of separable verb	0 (193)	16 (22)		16 (21)

with BCR. The forward pointing curved arrow emanating from the terminal node *sollen* 'should' indicates that the posterior clause is lacking its auxiliary and borrows it from the anterior clause. The backward pointing arrow is the secondary edge that denotes borrowing of the auxiliary *werden* 'be' by the anterior clause. Notice that secondary edges do not indicate the position of the borrowed constituent in the borrowing clause.

3.2 A methodological issue: Coordinate ellipsis vs. plausible conceptual inference

Figure 2 depicts FCR in sentence (10), which embodies a problematic aspect of the annotation in terms of secondary edges.

(10) *Wir verlieren sonst unser Profil und*
we lose otherwise our profile and
landen bei den Wahlen 1998 im Keller.
end-up at the elections 1998 in-the cellar
'Otherwise, we lose our profile and end up in the cellar at the 1998 elections'

In FCR, borrowing is restricted to left-peripheral major constituents of the anterior clause (see the FCR borrowing rule in Section 2). In (10), the left periphery only includes the subject NP *wir* because the conjuncts start to deviate already at the position of the finite verbs (*verlieren* 'lose' versus *landen* 'end up'). Hence, borrowing of the post-verbal modifier *sonst* 'otherwise' seems to violate the FCR borrowing rule. However, borrowing should be distinguished from PLAUSIBLE CONCEPTUAL INFERENCE. The fact that readers of sentence (10) tend to interpret *sonst* as modifying the posterior conjunct, is based on semantic/pragmatic knowledge rather than on knowledge of syntax. There are no SYNTACTIC reasons to include *sonst* as part of the posterior conjunct: Without this modifier, the conjunct would not be ungrammatical. In contrast, the inclusion of *wir* IS needed to complete the clause headed by *landen*: Without a subject NP, this active finite clause would be ill-formed.

This calls for an evaluation of the status of secondary edges: If the syntactic well-formedness of a conjunct is not affected by removing such an edge, we consider it a case of plausible conceptual inference rather than borrowing licensed by coordinate ellipsis. (This holds for the borrowing of *sonst* in (10).) Only if removal of the edge would make the conjunct ungrammatical (e.g., due to incompleteness of the subcategorization frame of a verb), we classify the edge as a case of genuine coordinate ellipsis (e.g., the borrowing of *wir* in (10)).

When classifying the secondary edges in each of the coordinate ellipsis types, we proceeded as follows.

GAPPING AND ITS SUBTYPES. The borrowing rule for these cases states that all non-contrastive major constituents are borrowed, except for negation elements (annotated by an NG edge). So we only needed to check whether the anterior clause included any non-contrastive major constituent that was not annotated as a secondary edge.

FCR. Left-peripheral borrowing of major constituents is mandatory here. Hence, in every FCR case, we determined the anterior clause's left-periphery, that is, the string from the leftmost major constituent up to and including the rightmost major constituent dominated by a secondary edge. If this string includes one or more major constituents without a secondary edge, this was counted as a potential violation of the borrowing rule. In Figure 2, the left-periphery consists of *wir verlieren sonst*, with *verlieren* indicating a potential borrowing violation. For all such patterns, we judged whether or not the secondary edges could denote plausible conceptual inferences. If so, the left periphery was readjusted by hand. For instance, as we judged *sonst* to be a plausible inference, the left periphery was reduced to *wir*, implying that the borrowing pattern in this sentence agrees with the rule.

BCR. For this ellipsis variant, we used the following definition of the right-periphery of the posterior clause: an uninterrupted string of major constituents dominated by secondary edges, extending backward from the end of the clause. We dealt with right-peripheral borrowings as if they were the mirror image of left-peripheral borrowing—though with an important exception: The leftmost constituent of the right-periphery need not be a complete major constituent (e.g. *o'clock* in (6)).

SGF. In addition to the subject NP, the posterior conjunct may only borrow—optionally—the clause-initial modifier of the anterior conjunct (e.g. *why* in (7)). So, the only possible violations of this rule are: borrowings of another type of major constituent, or of only a fragment of the clause-initial adverbial modifier, or of a constituent located to the right of the head verb. In such cases, we judged

whether the corresponding secondary edge could be based on plausible conceptual inference rather than coordinate ellipsis.

We realize that the distinction between two types of secondary edges as well as the criteria we used to classify them, are 'friendly' to the rather strict intuition-based borrowing rules put forward in Section 2. The annotators seem to have made their secondary-edge decisions on the basis of a much more liberal borrowing regimen. However, we reasoned it is good methodology to start from a more restrictive, more parsimonious theory and to adopt a less parsimonious one only after the more restrictive theory has been falsified.

4 Results

As can be gleaned from Table 1 in Section 2, TIGER contains 7194 sentences that include at least one clausal coordination, and 4046 of them have been annotated with one or more secondary edges in coordinated clauses. We classified each of these edges as representing genuine coordinate borrowings or plausible inferences. In the course of this process, we removed 26 sentences, chiefly for one of two reasons: The sentence includes an annotation error, or all of its secondary edges were deemed to represent plausible conceptual inference rather than ellipsis. The remaining 4020 TIGER sentences exhibit at least one exemplar of a genuine coordinate elliptical clausal structure. Actually, all seven main and subtypes of coordinate ellipsis are represented in the corpus. See the first row of Table 2 for the number of sentences exhibiting one of the four main ellipsis types.

We used the set of 4020 sentences to try and answer the following two questions:

- How accurately do the borrowing rules postulated in linguistic grammars—and used in computational parsers and generators—mirror the borrowing patterns observable in real texts? (In the absence of a treebank for spoken corpora, our answer will be restricted to *written* texts.)
- How can the frequencies of the various borrowing patterns help parsers to reconstruct borrowed (elided) constituents more accurately, and generators to produce more natural sounding and more easily interpretable coordinations of elliptical clauses?

These questions are discussed in separate Sections.

4.1 Correctness of the borrowing rules

After removing secondary edges that we judged to represent plausible conceptual inference, and readjusting left- or right-peripheries, we observed that in about 99 percent of the sentences the borrowing patterns agree with the intuition-based rules. Hence, we may conclude that these rules are not far off the mark. Nevertheless, we spotted some 40 sentences that violate a borrowing rule but, according to our judgment, are at least marginally acceptable. We discovered four borrowing (elision) patterns that may be characterized as 'fringe deviations' from the intuition-based coordinate ellipsis rules. Each of the offending patterns that we report here, is embodied in several sentences, hence is unlikely to reflect bad writing or sloppy editing.

OVERREDUCTION: In Gapping, FCR or SGF, only part of a major constituent is elided. In examples (11) and (12), both combining Gapping with BCR, the head noun of one remnant (of the subject of the posterior conjunct) is elided (indicated by strikethroughs). Furthermore, TIGER includes at least four sentences where the head of the PP is missing from the posterior conjunct. In (13), this holds for *aus* 'from'.

(11) *...während bei der Sparkasse X Gebühren von 50 und bei der Bank Y sogar ~~Gebühren~~ von 60 Mark zu berappen sind*
'... whereas at Savings Bank X fees of 50 and at Bank Y even ~~fees~~ of 60 Mark have to be coughed up'

(12) *Dabei schrumpfte der Auftragseingang aus dem Inland um drei und ~~der Auftragseingang~~ aus dem Ausland um vier Prozent*
'Moreover, the number of domestic orders shrank with three and ~~the number of orders~~ from abroad with four percent'

(13) *Das Anzeigengeschäft trug dazu 36 Prozent bei, aus dem Vertrieb kamen 34 Prozent und ~~aus~~ dem Druck 21 Prozent herein*
'The Advertising Department contributed 36 percent, 34 percent came in from Sales and 21 percent from Printing'

PERIPHERALITY VIOLATIONS BY LITTLE WORDS. In at least 10 FCR sentences, the third-person reflexive pronoun *sich* ('himself, herself, themselves') is located within the left-periphery of the anterior conjunct. In (14), *sich* is 'too late' to be shared by the other conjunct. (The end of the left-

peripheral region is indicated by slashes "//".) In (15), it is 'too early': It could be shared by the second conjunct, which however cannot use a reflexive pronoun. We also found a comparable case with pronominal NP *dies* 'this' and one with *niemals* 'never'.

The treebank contains one analogous example with BCR. In (16), the particle *an* causes a right-peripherality violation. The finite verb *berechnen* 'compute' is not a separable verb and does not have *an* as particle. However, it does need a direct object. This is elided here due to BCR, although its counterpart in the posterior clause is not right-peripheral.

(14) *... während // 78 Prozent sich für Bush und vier Prozent für Clinton aussprachen*
'... while 78 percent expressed themselves in favor of Bush and four percent for Clinton'

(15) *... daß sich weiß // davon am besten abhebt und von den Autofahrern am ehesten gesehen wird*
'... that [the color] white gives the better contrast and can be seen faster by the drivers'

(16) *[Sensoren ...] berechnen ~~die neue Position im Media-Land~~ und zeigen die neue Position im Media-Land an*
'[Sensors ...] compute and indicate the new position in Media-Land'

PERIPHERALITY VIOLATIONS BY CONTENT WORDS OR WORD GROUPS. In three sentences, a peripherality rule was violated by a content word, a word group, or even an entire subordinate clause. In FCR example (17), the posterior clause *noch immer gewährt* 'is still granting' borrows the direct object NP *Unterschlupf* 'shelter', implying that the left periphery is located to its right. This entails borrowing of PP *in der Vergangenheit* 'in the past', which however is semantically incompatible with the present tense of *gewährt* 'is granting'. In BCR sentence (18), the direct object NP *keine Garantie ...* 'no guarantee ...' borrowed by the anterior conjunct is not right-peripheral in the posterior conjunct but is followed there by the main verb *geben* and a complete extraposed complement clause. In BCR case (19), the passive auxiliary verb *werden* 'be' in the anterior conjunct is missing although a long extraposed PP follows its posterior counterpart. In TIGER, there are at least six BCR cases of the latter type (an extraposed constituent rightward of the presumed right periphery).

(17) *... das in der Vergangenheit so blutrünstigen Figuren [...] Unterschlupf // gegeben hatte bzw. noch immer gewährt*
'... which in the past had given shelter to bloodthirsty characters [...], resp. is still granting it'

(18) *Es gibt ~~keine Garantie dagegen daß [...]~~ und kann keine Garantie dagegen geben, daß [...]*
'There is ~~no guarantee~~ and there can be no guarantee that ...'

(19) *Nach und nach sollen dann auch Werke von exilierten Komponisten einbezogen ~~werden~~, der Aktionsradius erweitert werden auf Komponisten, die ...*
'By and by, works by exiled composers should be included, the radius of action extended to composers who ...'

SLOPPY GAPPING: remnants fulfilling a different grammatical function in the posterior conjunct than their counterpart in the anterior conjunct.[4] We found five cases (some perhaps intended as puns):

(20) *Es brachte [den SPD-Wirtschaftssprecher]$_{direct\ object}$ [um seinen Job]$_{modifier}$ und [der Öffentlichkeit]$_{indirect\ object}$ [eine heftige Debatte]$_{direct\ object}$*
'It cost the SPD speaker for economy his job and brought the public a severe debate'

(21) *Auwälder dienen [dem Hochwasserschutz]$_{indirect\ object}$ und [als Dschungel-Ersatz]$_{modifier}$*
'Riverside forests serve as protection against flooding and as jungle surrogate'

(22) Die Prinzessin erzählt im Fernsehen *[ihre Befindlichkeit]$_{direct\ object}$* und vielleicht auch *[von Männern]$_{modifier}$*
'On TV, the princess talks her sensitivities, and maybe also about men'

(23) *...1946 wurde er [Leiter ...]$_{predicate}$ und [mit ... betraut]$_{complement}$*
'In 1946 he became head and was entrusted with ...'

[4] Sentences (20) through (24) cannot be analyzed as (non-clausal) 'coordinations of unlikes'. In such coordinations (a famous English example being *John is a Republican and proud of it*), the conjuncts are 'unlike' in that they embody constituents of different categories (NP and AP in the example). However, the unlike conjuncts should be adjacent AND fulfill the same grammatical function. This combination of criteria is not met by sentences (20) through (24).

(24) ... sie ... ziehen *[Grimassen]*$_{direct\ object}$ und *[an den erkalteten Zigarillos]*$_{modifier}$ ...
'... they make grimaces and draw on the dead cigarillos'

To conclude, although nearly all clausal elliptical coordinations obey the borrowing rules, the four groups of fringe deviations call for some relaxation.

4.2 Implications for grammar, parsing, and generation

An improved grammar rule for BCR seems to require a more general definition of 'end of clause': A clause ends not only after its last word but also at the position that serves as a receptacle for extraposed constituents (i.e. just before the word *yesterday* in (6)). Sentences (17) through (19) would be ruled in by this amendment. The borrowing rules for FCR and BCR may be allowed to overlook little words such as personal and reflexive pronouns, and verb particles. The other TIGER sentences cited in Section 4.1, however, seem to require more subtle finetuning.

Table 2 shows the borrowing (elision) frequencies of various grammatical functions in the four main types of clausal coordinate ellipsis. For example, the constituents most likely elided in FCR are the subject and the complementizer. This frequency information can help a chunker or shallow parser to reconstruct elided elements and thus to recover from parsing failure. This presupposes, of course, that the analyzer has been able to recognize the clausal coordination and the ellipsis type. Given the success of the strongly peripherality-oriented borrowing rules (Kempen, in press), they provide a sound basis for the design of efficient parsers for coordinate structures.

As for generation, elliptical coordinations figure prominently in several application domains, e.g. weather forecasting. More concise and more variegated texts can be produced if the generator is able to apply the various types of elision to non-reduced sentences which express the intended meaning (Harbusch & Kempen, 2006). In many sentences, for instance, Gapping and BCR form competing but mutually exclusive ways of avoiding unnecessary reduplication of sentence fragments. Frequency data such as those in Table 2 can help to select a natural sounding elision option.

5 Future work

Our evaluation study with the TIGER treebank revealed that the intuition-based borrowing (elision) rules summarized in Section 2 cover about 99 percent of the corpus sentences. One of our goals is to build an efficient parser that heavily relies on these rules in its treatment of coordinate structures. Another future project is to elicit from native speakers of German grammaticality judgments for sentences that embody the fringe deviations we discovered and reported in Section 4.1. The results will hopefully serve to finetune the borrowing rules.

References

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith & Hans Uszkoreit (2004). TIGER: Linguistic Interpretation of a German Corpus. *Research on Language and Computation, 2,* 597-620.

Karin Harbusch & Gerard Kempen (2006). ELLEIPO: A module that computes coordinative ellipsis for language generators that don't. In: *EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics* (Trento, Italy; April 2006).

Janne B. Johannessen (1998). *Coordination*. Oxford: Oxford University Press.

Gerard Kempen (in press). Clausal coordination and coordinate ellipsis in a model of the speaker. *Linguistics*.

Esther König & Wolfgang Lezius (2003). The TIGER language – A Description Language for Syntax Graphs: Formal Definition. Tech. Rep. IMS, University of Stuttgart.

Ivan A. Sag, Thomas Wasow & Emily M. Bender (2003). Syntactic Theory: A formal introduction, Stanford: CSLI Publications, Second Edition.

Mark Steedman (2000). *The syntactic process.* Cambridge MA: MIT Press.

John R. te Velde (2006). *Deriving Coordinate Symmetries: A phase-based approach integrating Select, Merge, Copy and Match*. Amsterdam: Benjamins.

Robert R. van Oirsow (1987). *The syntax of coordination*. London: Croom Helm.