Generating clausal coordinate ellipsis multilingually: A uniform approach based on postediting

Karin Harbusch
Computer Science Department
University of Koblenz-Landau
PO Box 201602, 56016 Koblenz/DE
harbusch@uni-koblenz.de

Gerard Kempen
Max Planck Institute for Psycholinguistics
PO Box 310, 6500AH Nijmegen/NL
& Cognitive Psychology Unit, Leiden University
gerard.kempen@mpi.nl

Abstract

Present-day sentence generators are often incapable of producing a wide variety of well-formed elliptical versions of coordinated clauses, in particular, of combined elliptical phenomena (Gapping, Forward and Backward Conjunction Reduction, etc.). The applicability of the various types of clausal coordinate ellipsis (CCE) presupposes detailed comparisons of the syntactic properties of the coordinated clauses. These nonlocal comparisons argue against approaches based on local rules that treat CCE structures as special cases of clausal coordination. We advocate an alternative approach where CCE rules take the form of postediting rules applicable to nonelliptical structures. The advantage is not only a higher level of modularity but also applicability to languages belonging to different language families. We describe a language-neutral module (called Elleipo; implemented in JAVA) that generates as output all major CCE versions of coordinated clauses. Elleipo takes as input linearly ordered nonelliptical coordinated clauses annotated with lexical identity and coreferentiality relationships between words and word groups in the conjuncts. We demonstrate the feasibility of a single set of postediting rules that attains multilingual coverage.

1 Introduction

In present-day Natural-Language Generation (NLG) architectures, elision rules typically form part of the *Aggregation* component, i.e. of the module that decides how to group conceptual messages into a sentence—a module belonging to the *Microplanner* (cf. Reiter & Dale, 2000, for an authoritative overview of sentence and text generation technology). In such generators, the computation of coordinate structures takes place at a relatively early stage of syntactic processing. However, many types of clausal coordinate ellipsis (CCE) require detailed comparisons of the final syntactic shape of the coordinated clauses (*conjuncts*). This is even more true when it is desirable to combine elision constructions, as in German example (1a), where Subgapping—a form of Gapping—combines with Backward Conjunction Reduction (for definitions and examples see Table 1). Example (1) also illustrates that often more than one elliptical option is available: (1b) shows a variant with Subgapping alone. If the nonelliptical sentence generator would choose a different Verb order in the second conjunct (‘*gestutzt werden sollen’* as in (1d)), then Subgapping would be the sole alternative.

(1) a. *Die Bäume sollen gefällt ~~werden~~ und*
The trees should cut-down be and
die Sträucher ~~sollen~~ gestutzt werden
the shrubs should pruned be
'The trees should be cut down and the shrubs pruned’
b. *Die Bäume sollen gefällt werden und die Sträucher ~~sollen~~ gestutzt ~~werden~~*
c. **Die Bäume sollen gefällt ~~werden~~ und gestutzt werden ~~sollen~~ die Sträucher*
d. *Die Bäume sollen gefällt werden und gestutzt ~~werden~~ ~~sollen~~ die Sträucher*

The comparisons between the clausal conjuncts mainly pertain to the linear order of their major constituents and to identity relations between the lexical material contained in them. For instance, if a right-peripheral string of lexical items in the anterior conjunct is identical to such a string in the posterior conjunct, then Backward Conjunction Reduction licenses elision of the former string. In (1a), the two one-word strings ‘*werden’* meet this requirement.

Language-typological work, e.g. the recent survey by Haspelmath (2007), provides another argument for a “late” CCE component. The main phenomena can be categorized into a small number of basic types, which have been attested in languages belonging to different language

families. This suggests the possibility of a multilingual approach to CCE where the main CCE processes are defined as procedures that are isolated from the "normal" grammar rules for nonelliptical structures, and are independent from each other (see Section 4).

Instead of having an aggregation component where the rules for nonelliptical clausal coordinate structures are intermingled with rules for elliptical variants, we consider an alternative approach where the application of the ellipsis rules is deferred until the nonelliptical structures have been completed. That is, the elision options are calculated and executed during a *postediting* stage, after the strategic and tactic components of the generator have delivered the nonelliptical versions. We claim that this modular approach facilitates the development of multilingual CCE components for different languages by switching on and off the individual CCE procedures (e.g. no Gapping in Japanese).

The structure of the paper is as follows. In Section 2, we introduce the four main CCE phenomena and informally define their treatment in a procedural manner. Section 3 lays out the basic postediting rules that we implemented in JAVA as a language-neutral algorithm (nicknamed *Elleipo*, which means 'I leave out' in classical Greek). In Section 4, we present some findings from language-typological studies and explore their implications for potential multilingual applicability of Elleipo. Finally, in section 5, we draw some conclusions.

The Elleipo version described here embodies several important improvements to a version described briefly in Harbusch & Kempen (2006), particularly with respect to *Subject Gap in clauses with Finite/Fronted Verbs (SGF)*. Moreover, the space allowed here enables us to explain Elleipo's inner workings in more detail, and to demonstrate its multilingual potential.

2 Clausal coordinate ellipsis (CCE)

2.1 Clausal coordinate ellipsis in linguistic theories and in NLG: State of the art

Treatments of the phenomena of clausal coordination and CCE are provided by all major grammar formalisms. Some representative studies are Sarkar & Joshi (1996) for Tree Adjoining Grammar; Steedman (2000) for Combinatory Categorial Grammar; Bresnan (2000) and Frank (2002) for Lexical-Functional Grammar; Crysmann (2003) and Beavers & Sag (2004) for Head-driven Phrase-Structure Grammar; and te Velde (2005) for the Minimalist Program. Their treatments of CCE take the form of special declarative coordination rules, in contrast with the modular and procedural approach we propose.

In the NLG community, modular treatments of CCE—implemented as programs that take unreduced coordinations expressed in the system's grammar formalism as input and return elliptical versions as output—have been elaborated in several projects (Shaw, 1998; Dalianis, 1999; Hielkema, 2005). These systems are limited in that that they do not cover all of the four CCE processes and are monolingual.

2.2 Clausal coordinate ellipsis types

In the linguistic literature on clausal coordinate ellipsis, four main CCE processes are often distinguished, as shown in Table 1.

In the theoretical framework by Kempen (2009) and its implementations (Harbusch & Kempen (2006) for German and Dutch; Harbusch *et al.* (2009) for Estonian), the elision process is guided by *identity constraints* and *linear order* (cf. column 4 in Table 1). We distinguish three basic types of identity relations between words or word groups (constituents) belonging to different conjuncts[1]:

(1) *Lemma identity:* two different words belong to the same inflectional paradigm; e.g. the Verbs '*live'* and '*lives'* in example (2).

(2) *Form identity:* two words have the same spelling/sound and are lemma-identical; e.g., two tokens of '*want'* are form-identical if they are both Verbs, but not if one is a Verb and the other is a Noun.

(3) *Coreferentiality:* two words/constituents denote the same entity or entities in the external context, i.e. have the same reference.

[1] Very often, lemma- and form identity coincide with coreference, but not necessarily. For instance, in '*John bought a car in July, and Peter ~~bought a car~~ in August*,' the two tokens of '*a car'* are not, in all likelihood, coreferential. Nevertheless, elision of '*a car'* is allowed in this Gapping example. In the semantic literature, this relation is called *sloppy identity*. On the other hand, in '*Who wants coffee and who wants tea?,'* the two tokens of '*who'* are not coreferential, and the second token cannot be elided. We assume that the strategic and/or the tactical component of the generator assigns differing identity tags (see Section 3.1) to lemma- or form-identical constituents if and only if their reference is strictly non-identical. Also note that, in the following, the three identity relationships will not only be applied to individual words but also to constituents entirely consisting of words that meet the respective criteria (cf. the numerical subscripts in Figure 1 in Section 3).

Table 1. Clausal coordinate ellipsis (CCE) types. Column 2 lists the abbreviations for the types mentioned in column 1 (see Elleipo's algorithm in Section 3). Column 3 illustrates the CCE types. Column 4 summarizes the elision conditions explained in Section 3.

CCE type	Abbr.	Examples	Elision conditions
Gapping	*g*	*(2) Ulf* ***lives*** *in Leipzig and his children* ~~***live***~~$_g$ *in Ulm*	Lemma identity of Verb & contrastiveness of remnants
Long-Distance Gapping (LDG)	*g(g)*$^+$	*(3) My wife* ***wants to buy*** *a car and my son* ~~***wants***~~$_g$ ~~***[to buy]***~~$_{gg}$ *a motorcycle*	Gapping conditions in *superclause* (Section 3.2.1)
Subgapping	*sg*	*(4) The driver* ***was*** *killed and the passengers* ~~***were***~~$_{sg}$ *severely wounded*	Gapping conditions & VP remnant in second conjunct
Stripping	*str*	*(5) My sister* ***lives in Narva*** *and her children* ~~***[live in Narva]***~~$_{str}$ ***too***	Gapping conditions & only one non-Verb remnant
Forward Conjunction Reduction (FCR)	*f*	(6) ***Since two years, my sister*** *lives in Delft and* ~~***[since two years, my sister]***~~$_f$ *works in Leiden* (7) *Tokyo is the city [S* ***where*** *Ota lives and* ~~***where***~~$_f$ *Kusuke works]*	Form identity & left-peripherality (within clause boundaries) of major clausal constituents
Backward Conjunction Reduction (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
Subject Gap in clauses with Finite/Fronted Verbs (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 4 of Table 1, all forms of *Gapping* (i.e. including *LDG, Subgapping* and *Stripping*) are characterized by elision of the posterior member of a paired lemma-identical Verb. The position of this Verb need not be peripheral but is often medial, as in examples (2) through (5), and (9). Non-elided constituents in the posterior conjunct are called *remnants*. All remnants should pair up with a constituent in the anterior conjunct that has the same grammatical function but is not coreferential. 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 two tokens of '*my'* in (3) occupy comparable positions in the two conjuncts, it is not possible to elide any of them. On the other hand, '*were'* in (4) can be elided from the posterior conjunct although it has no literal anterior counterpart.)

In *LDG*, the remnants originate from different clauses (more precisely: different clauses that belong to the same *superclause*; term defined in Section 3.2.1). 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 the Adverb '*too.'*

In *Forward Conjunction Reduction (FCR)*, elision affects the posterior token of a pair of left-peripheral strings consisting of one or more form-identical major constituents. In (6) and (7), the posterior tokens of '*since two years, my sister'* and '*where,'* respectively, belong to such pairs and are eligible for FCR.

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 ('*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—witness examples like (8).[2]

Subject Gap in clauses with Finite/Fronted Verbs (SGF) can elide the Subject of the posterior conjunct when in the anterior conjunct the form-identical Subject follows the Verb (Subject-Verb inversion); moreover, the Head Verbs of the conjoined clauses—both with main or interrogative clause word order—are different. (FCR cannot have caused the absence of the posterior Subject since the anterior Subject is not left-peripheral.) The examples in (11)

[2]However, case-identity is required as well, at least in German: *?Hilf* ~~*[dem Patienten]*~~$_{DAT}$ *und reanimier [den Patient]*$_{ACC}$ 'Help and reanimate the patient'.

through (14) show that the first constituents of the unreduced clauses must meet certain special requirements, which extend the rule proposed in our previous publications. In particular, these constituents *are* allowed to be non-form-identical finite Head Verbs (11) or form-identical Modifiers (12) but *not* form-identical arguments, e.g. Direct Objects (13) or Complements (14). Additionally, if FCR is licensed, as in (12), it should actually be realized in order to allow SGF.

(11) *Stehen die Leute noch am Eingang und*
Stand the people still at-the entrance and
rufen ~~*[die Leute]*~~$_s$ *Parolen?*
shout the people slogans
'Are the people still standing at the entrance (and are they) shouting slogans?'

(12) *Warum/Gestern bist du gegangen und*
Why/Yesterday have you left and
[~~*warum/gestern*~~*]*$_f$ *hast* ~~*du*~~$_s$ *nichts gesagt*
why/yesterday have you nothing said
'Why did you leave and didn't you tell me anything?' / 'Yesterday you left and ...'

(13) **Diesen Wein trinke ich nicht mehr und*
This wine drink I not anymore and
[~~*diesen Wein*~~*]*$_f$ *gieße* ~~*ich*~~$_s$ *weg*
this wine throw I away
'I don't drink this wine anymore and throw it away'

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

3 Language-neutral CCE generation

In this Section, we describe Elleipo's algorithm in more detail than we were able to in Harbusch & Kempen (2006), again using the German example (15). Moreover, we elaborate on SGF, given the new, more detailed rules. We limit ourselves to '*and'*-coordinations of only n=2 conjuncts. Actually, Elleipo can handle n-ary coordinations consisting of $n \geq 2$ conjuncts by processing n–1 consecutive pairs of conjuncts (1+2, 2+3, etc.), together with an asyndeton rule that replaces non-final *'and'*-s by commas.

(15) *Heute wird Hans sein Auto putzen und*
Today will Hans his car clean and
~~*heute wird*~~ *Susi ihr Fahrrad* ~~*putzen*~~
today will Susi her bike clean
'Today, Hans will clean his car and today, Susi will clean her bike'

Elleipo's functioning is based on the assumption that CCE does not result from the application of local declarative grammar rules for clause formation but from a procedural component that inspects nonelliptical (unreduced) sentences produced by the sentence generator and may block the overt expression of certain constituents. Due to this feature, Elleipo can be combined, at least in principle, with various generators. However, the module needs a formalism-dependent interface that converts generator output to a (simple) canonical form.

3.1 Elleipo's input

Elleipo takes as input nonelliptical syntactic trees in *canonical form*, supplied with *identity tags* (cf. Figure 1). Every categorial node of an input tree is immediately dominated by a functional node. Each conjunct is rooted in a categorial node whose daughter nodes (immediate constituents) are grammatical functions (Subject, Direct Object, Head, Subordinating Conjunction, Expr(ession), etc.). Within a conjunct, all major constituents are represented at the same hierarchical level ("flat" trees).

Categorial nodes are adorned with numerical identity tags (ID-tags) which express lemma identity. In Figure 1, the ID "2" is attached to the head node of both exemplars of AP '*heute'* 'today', thus marking their lemma identity. In contrast, the Subject NPs '*Hans*' and '*Susi*' carry different ID-tags, indicating that they are not lemma-identical and cannot be elided by any CCE process.

3.2 The three stages of Elleipo

Elleipo is called for every coordination domain within a non-elliptical input clause. We define a *coordination domain* as a (sub)tree rooting in a grammatical function node that dominates two or more categorial S-nodes separated by coordinating conjunctions ('*and'*). For any given coordination domain, Elleipo's task consists of three consecutive stages: *Preparation, Diagnosis, and ReadOut.*

3.2.1 Preparation

The first job within *Preparation* is the demarcation of *superclauses*. Kempen (2009) introduced this notion in the treatment of Gapping, in particular LDG. A *superclause* is either a simple finite or non-finite clause (rooting in an S-node, without any subordinate clauses), or a hierarchy of finite or non-finite clauses where every embedded clause is an immediate daughter of an embedding clause; moreover, none of the participating clauses begins with a subordinating

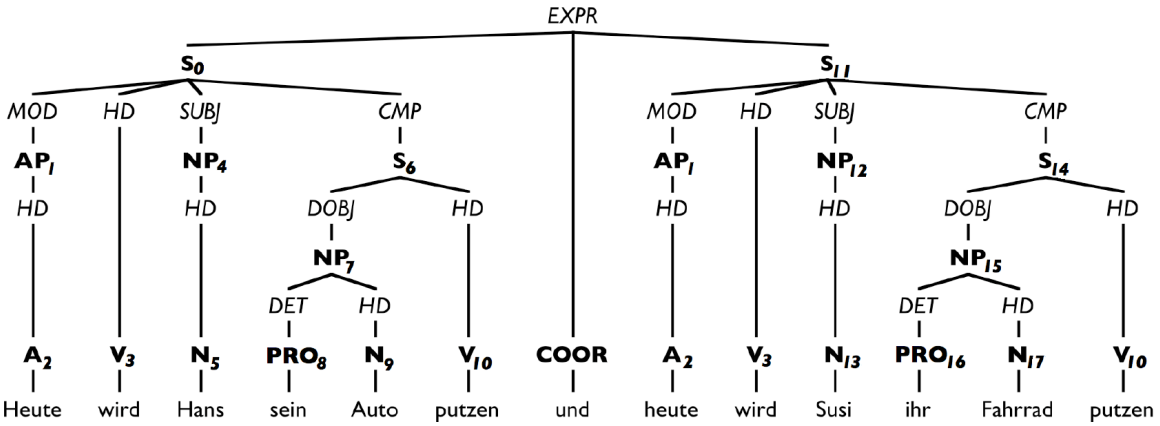

Figure 1. Non-elliptical input tree in canonical form, underlying sentence (15). Categorial nodes are printed in bold, functional nodes in italics. The numerical subscripted tags denote lemma identity or coreference.

conjunction, with the possible exception of the topmost member of the hierarchy.[3]

Next, Elleipo inspects and compares the content of the conjoined clauses by assembling four lists: FUNC-PAIRS, LI-FUNC-PAIRS, LPERIPH and RPERIPH (see Table 2). The lists FUNC-PAIRS and LI-FUNC-PAIRS are crucial not only in calculating whether a form of Gapping is applicable but also in the determination of *contrastiveness* of Gapping remnants. We presuppose a division of MODifier constituents into MOD types—locative (LMOD), temporal (TMOD, causal (CMOD), etc.—which are recorded in the two lists of pairs. Gapping requires the set of grammatical functions, including MOD types, in the anterior and posterior conjuncts to be identical. If so, and if FUNC-PAIRS includes at least one pair of non-coreferential members (carrying different ID-tags), the Boolean variable CONTRAST is set to *true*. In the example, FUNC-PAIRS(S_0, S_{11}) includes two pairs of non-coreferential major constituents (the Subjects and the Complements); hence, CONTRAST = *true*. LPERIPH is crucial in FCR, where only complete form-identical major constituents may be elided. RPERIPH is used in BCR, which sometimes leaves incomplete constituents behind, as exemplified by (8) and (9).

[3] The "embedded" clauses referred to in the definition of superclause fulfill the grammatical function of Subject or Object Complement within the embedding clause, or they are adverbial clauses fulfilling the function of Modifier within the embedding clause. In Figure 1, the Complement clauses S_6 and S_{14} are major constituents of S_0 and S_{11}, respectively. The hierarchy spanning S_0 and S_6 is a superclause, and so is the hierarchy consisting of S_{11} and S_{14}. In '*Hans sagte, dass Susi ihr Fahrrad putzen wird'* 'Hans said that Susi will clean her bike,' the Complement clause does not belong to the same superclause as the main clause '*Hans sagte* ...,' but instead starts up its own superclause. Gapping and its varieties can only be applied to two coordinated superclauses.

3.2.2 Diagnosis

For each of the four CCE processes, Elleipo inspects all coordination domains for elision options. This requires interpreting the lists collected during the Preparation stage. Any licensed elision option for a word or constituent causes the current value of the parameter CCE-TYPE to be added as a tag to that word or constituent (cf. the subscripts in examples (2) through (10)). Different elliptical variants (cf. examples (1a/b)) are represented by multiple tags and yield alternative realizations, to be spelled out during the final ReadOut stage. If the Boolean variable CONTRAST is *true*, Gapping runs recursively within the current coordination domain. Figure 2 shows pseudocode for Gapping, with input parameters LC = left conjunct, RC=right conjunct, and CCE-TYPE=*g*). In the example: GAPPING(S_0, S_{11}, "*g*").

Lemma identity of the Head Verbs of the clausal conjuncts licenses Gapping. So, the temporal Adverbial Modifier and the Head Verb of the posterior conjunct can both be marked for elision: '*Heute wird Hans sein Auto putzen und* ~~*heute*~~$_g$ ~~*wird*~~$_g$ *Susi ihr Fahrrad putzen'* (steps 8 and 9). Earlier on, in step 3 and 4, Elleipo has already noticed that one of the non-lemma-identical pairs—S_6 and S_{14}—consists of Complement clauses belonging to the same superclause as the coordinated main clauses (i.e. they do not start up a new superclause hierarchy). In step 6, Elleipo is called recursively for this coordinate subdomain, with argument CCE-TYPE set to "*gg*". As the Head Verbs of these complement clauses are lemma-identical and the contrastiveness condition holds (i.e. the grammatical function DOBJ occurs in both the anterior and the posterior complement and the exemplars are not lemma-identical), the posterior Verb is marked for elision, yielding '*Heute wird*

Table 2. Definitions of the (possibly empty) lists of paired major clause constituents calculated during Elleipo's Preparation stage. Column 3 shows the content of the lists for example (15), i.e. the superclauses S_0 and S_{11}.

List	Definition	Content of lists for example (15)
FUNC-PAIRS	All constituent pairs *LCAT–RCAT* with same grammatical function, dominated by an S-node pair; if (LCAT, RCAT) is an S-node pair, then FUNC-PAIRS is assembled recursively for this pair as well.	FUNC-PAIRS(S_0,S_{11})={AP_1–AP_1,V_3–V_3, NP_4–NP_{12}, S_6–S_{14}} Due to recursive application: FUNC-PAIRS(S_6,S_{14})={NP_7–NP_{15},V_{10}–V_{10}}
LI-FUNC-PAIRS	Lemma-Identical pairs of corresponding FUNC-PAIRS (i.e., LI-FUNC-PAIRS ⊆ FUNC-PAIRS).	LI-FUNC-PAIRS(S_0,S_{11})={AP_1–AP_1,V_3–V_3} LI-FUNC-PAIRS(S_6,S_{14})={V_{10}–V_{10}}
LPERIPH	Left-peripheral form-identical complete major constituents shared by the conjuncts.	LPERIPH(S_0,S_{11})={A_2,V_3}
RPERIPH	Right-peripheral lemma-identical lexical string shared by the conjuncts.	RPERIPH(S_0,S_{11}) ={V_{10}}

```
1   proc GAPPING(LC, RC, CCE-TYPE) {
2    for all pairs (LCAT, RCAT) in FUNC-PAIRS(LC, RC) {
3     if (LCAT is an S-node) & (LCAT doesn't begin a new superclause) then {// call GAPPING recursively//
4        if NOT (LCAT and RCAT are lemma-identical)
5           then {attach "g" to CCE-TYPE; //LDG//
6                 call GAPPING(LCAT, RCAT, CCE-TYPE);}
7           else mark RCAT for elision, with CCE-TYPE}
8     if (LCAT and RCAT are lemma-identical) & NOT(LCAT is an S-node)
9       then mark RCAT for elision, with CCE-TYPE}}
```

Figure 2. Pseudocode for the GAPPING procedure

Hans sein Auto putzen und ~~heute~~$_g$ ~~wird~~$_g$ Susi ihr Fahrrad ~~putzen~~$_{gg}$.'

FCR and BCR are both executed by one procedure, called CR. In FCR mode, CR is called with the value of LPERIPH as input; in BCR mode, it takes RPERIPH's value as input. Recall that these lists were computed in the Preparation stage and may contain a form-identical (LPERIPH) or a lemma-identical (RPERIPH) lexical string. In calls to CR (see the pseudocode in Figure 3), parameter PERIPH is set to LPERIPH or RPERIPH, and *CCE-TYPE* to "*b*" or "*f*" depending on whether BCR or FCR is to be executed. In our example, the main program calls are:

CR(S_0, S_{11}, LPERIPH(S_0, S_{11}), "*f*"), and
CR(S_0, S_{11}, RPERIPH(S_0, S_{11}), "*b*").

FCR and BCR are attempted after, and fully independently from, Gapping, irrespective of whether the latter was successful or not. As Modifier '*heute'* and Head Verb '*wird'* are both listed in *LPERIPH*(S_0, S_{11}), both major constituents are marked as eligible for elision from the posterior conjunct (line 6 of Figure 3)—an effect which happens to coincide with the effects of Gapping:

'*Heute wird Hans sein Auto putzen und ~~heute~~$_{g\text{-}f}$ ~~wird~~$_{g\text{-}f}$ Susi ihr Fahrrad ~~putzen~~$_{gg}$'.*

```
1   proc CR(LC, RC, PERIPH, CCE-TYPE) {
2   while PERIPH ≠ ∅ {
3      set (LCAT, RCAT) to PERIPH's first element;
4      PERIPH =PERIPH minus first element;
5      if CCE-TYPE = "f"
6         then mark RCAT else LCAT for elision,
                with CCE-TYPE}}
```

Figure 3. Pseudocode for procedure CR (executing FCR or BCR).

When attempting BCR, Elleipo discovers the lemma-identical '*putzen'* (V_{10}), and marks the anterior exemplar with "*b*": '*Heute wird Hans sein Auto ~~putzen~~$_b$ und ~~heute~~$_{g\text{-}f}$ ~~wird~~$_{g\text{-}f}$ Susi ihr Fahrrad ~~putzen~~$_{gg}$'.*

Elleipo's fourth check concerns SGF (Figure 4; see section 2.1 for the rules), here with negative result. In example (15), Subject-Verb-inversion is realized in the first conjunct. However, the two Subjects *'Peter'* and *'Susi'* are not coreferential.

```
1   proc SGF(LC, RC) {
2     if (Head Verb precedes SUBJ in LC)
        & (coreferential SUBJs in LI-FUNC-PAIRS)
        & (Head Verb or MOD in 1st position in LC)
        & (1st position in RC is occupied by SUBJ or a
            major constituent already marked for FCR)
3       then mark RC's SUBJ for elision "s";}
```

Figure 4. Pseudocode for SGF

3.2.3 ReadOut

The resulting terminal string annotated with elision marks is handed over to the *ReadOut* stage. As ReadOut assumes that all elisions are optional, it may deliver more than one elliptical output string. However, not every possible combination of elisions is legitimate; certain combinations have to be ruled out. We mention four important restrictions here. First, Gapping and BCR cannot elide both tokens of a lexical item. For instance, if '*putzen'* in the anterior conjunct of (15) is elided due to BCR, then its posterior counterpart '*putzen,'* which could be Gapped, should remain—and vice-versa. Second, in LDG, if a Verb with *n* subscripts "*g*" is elided, then all Verbs with *m*>*n* subscripts "*g*" should be elided as well. Third, in Gapping, if only one non-Head-Verb constituent remains (i.e. Stripping), then (the language-specific equivalent of) the Adverb '*too'* is added. Fourth, SGF requires that FCR, if licensed, is actually executed. Moreover, the ReadOut stage performs certain types of embellishments, e.g. it applies an asyndeton rule that replaces all but the last token of the coordinating conjunction by commas.

3.4 Elleipo evaluated for German

A detailed evaluation of Elleipo is currently only available for the German version (Harbusch & Kempen, 2007). In the TIGER corpus with 50,000 sentences, 99 percent of the CCE sentences conform to Elleipo's CCE rules. Nevertheless, we are aware that these rules do not handle SGF in conjoined subordinate clauses where the first conjunct has the standard Verb-final word order but the second conjunct (with SGF) embodies Verb-second order. Furthermore, Elleipo does not take into account certain semantic constraints ("one-event semantics"; Reich, in press; see also Frank, 2002; Kempen, 2009). Another insufficiency concerns the rules for asyndeton, which are more complicated than simply converting prefinal *'and'*-s to commas (see Borsley (2005) for pertinent examples).

4 Multilingual CCE generation

4.1 CCE rules in typological studies

The four CCE processes have been attested in two Germanic languages (German and Dutch) and in a Finno-Ugric language (Estonian; Harbusch *et al.*, 2009), where they account for a wide range of CCE phenomena. This invites the prediction that CCE obeys the same rules in many other languages as well. However, Haspelmath's (2007) survey immediately falsifies this prediction: Other CCE processes may be at work in other languages, and/or some of the above four main processes may be absent.

Japanese may provide illustrations of both points. On the one hand, it is uncontroversial that it does not have Gapping. On the other hand, it may have a form of CCE that stands midway between FCR and BCR. Yatabe (2001) interprets (16) as *Left Node Raising*, i.e. as the mirror image of BCR. Like FCR, it elides a left-peripheral string of the posterior conjunct; like BCR, the elided string need not be a complete major constituent. The elided Verb *yonde* is part of the prenominal Relative clause *yonde agenakatta* which is a major constituent (immediate daughter) of the NP headed by the Noun *hito*. But notice that (16) embodies coordination of NPs rather than clauses. If Japanese indeed exhibits *partial* elision of left-peripheral major constituents at the *clausal* level, thus violating our FCR definition, then we obviously need to define an additional CCE type.

(16) *Yonde ageta hito to*
read$_{gerund}$ give$_{past}$ person and
~~*yonde*~~$_f$ *agenakatta hito ga ita*
read$_{gerund}$ give$_{neg\text{-}past}$ person NOM be$_{past}$
'There were people who gave (him/her) the favor of reading (it) (to him/her) and people who didn't'

In contrast, Abe & Hoshi (1997) analyze Japanese example (17) in terms of *Preposition Stranding*. As far as we can see, this structure does not require a special CCE process because Elleipo treats it as BCR, which allows partial elision of the PP Modifier in the anterior conjunct, hence stranding of the Preposition.

(17) *John-ga Bill[-*~~*nituite hanasita*~~*]*$_b$*, sosite*
John-Nom Bill -about talked and
Mary-ga Susan-nituite hanasita
Mary-Nom Susan-about talked
'John talked about Bill and Mary about Susan'

Haspelmath (2007) also discusses certain languages with Subject–Object–Verb (SOV) as basic word order (Turkish and Korean) which allow Object deletion from non-peripheral positions in the posterior conjunct; i.e., they license the pattern SOV&S_V, as in '*The-boy the-cart pulled and the-girl* ~~*the cart*~~ *pushed*.' Elleipo cannot handle

this CCE structure: FCR and BCR require peripherality of the elided constituent; SGF only applies to Subjects; and Gapping presupposes elision of the Head Verb. In order to encompass the problematic pattern, we may need to define a new CCE process. However, at least in Turkish SOV&S_V cases, the elision may be due to pragmatic factors. Göksel & Kerslake (2005) show that major clause constituents fulfilling diverse grammatical functions can be elided as long their referents are recoverable on the basis of the accompanying linguistic or nonlinguistic context. Because the anterior conjunct may provide such a context, one first needs to rule out contextual recoverability as the licensing factor.

At the same time, Haspelmath also shows that Elleipo's four CCE processes cover a high proportion of CCE patterns occurring cross-linguistically. (However, he does not discuss SGF.) A typical illustration is the set of nine "more widely attested patterns" of CCE that he enumerates with respect to elision of Objects or Verbs in four language groups with different basic word orders of S, O and V (Table 2 in Haspelmath, 2007). All these patterns are covered by our four CCE processes, except SOV&S_V.

5 Discussion

We conclude that a software module embodying Elleipo's four main CCE processes—maybe with relatively minor adjustments—will be able to generate a great deal of CCE structures for many different languages.

As for possible practical applications, Elleipo's status as a postprocessor working on input specifications of unreduced syntactic structures facilitates combinability with sentence generators based on various grammar formalisms. Even template-based message generators, such as used in car navigation and weather forecast systems, can attain higher levels of fluency and conciseness if the templates are annotated with syntactic structure and ID-tags.

References

Abe, J. & Hoshi, H. 1997. Gapping and P-Stranding. *Journal of East Asian Linguistics, 6*.

Beavers, J. & Sag, I.A. 2004. Coordinate Ellipsis and Apparent Non-Constituent Coordination *Procs. of 11th Int. Conf. on HPSG*, Louvain.

Borsley, R.D. 2005. Against ConjP. *Lingua, 115*.

Bresnan, J.W. 2000. *Lexical-Functional Syntax.* Blackwell, Oxford, UK.

Crysmann, B. 2003. An asymmetric theory of peripheral sharing in HPSG: Conjunction reduction and coordination of unlikes. *Procs. of 8th Conf. on Formal Grammar*, Vienna.

Dalianis, H. 1999. Aggregation in natural language generation. *Computational Intelligence, 15*.

Göksel, A. & Kerslake, C. 2005. *Turkish: A comprehensive Grammar.* Routledge, Abington, Oxon, UK.

Frank, A. 2002. A (discourse) functional analysis of asymmetric coordination. *Procs. of LFG02 Conf.*, Athens.

Harbusch, K. & Kempen, G. 2006. Elleipo: A module that computes coordinative ellipsis for language generators that don't. *Procs. of 11th EACL,* Trento.

Harbusch, K. & Kempen, G. 2007. Clausal coordinate ellipsis in German. *Procs. of 16th NODALIDA*, Tartu.

Harbusch, K., Koit, M. & Õim, H. 2009. A comparison of clausal coordinate ellipsis in Estonian and German. *Procs. of 12th EACL*, Athens.

Hielkema, F. 2005. *Performing syntactic aggregation using discourse structures.* Unpublished Master's thesis, AI Unit, University of Groningen.

Haspelmath, M. 2007. Coordination. In: Shopen, T. (Ed.), *Language typology and linguistic description.* Cambridge University Press [2nd Ed.], Cambidge UK.

Kempen, G. 2009. Clausal coordination and coordinate ellipsis in a model of the speaker. *Linguistics*, *47(3)*.

Reich, I. In press. From discourse to "odd coordinations"—On Asymmetric Coordination and Subject Gaps in German. In: Fabricius-Hansen, C. & Ramm, W. (Eds.), *'Subordination' vs. 'Coordination' in Sentence and Text*. Benjamins, Amsterdam.

Reiter, E. & Dale, R. 2000. *Building natural language generation systems.* Cambridge University Press, Cambridge, UK.

Sarkar, A. & Joshi, A. 1996. Coordination in Tree Adjoining Grammars. *Procs. of 16th COLING,* Copenhagen.

Shaw, S. 1998. Segregatory coordination and ellipsis in text generation. *Procs. of 17th COLING,* Montreal.

Steedman, M. 2000. *The syntactic process.* MIT Press, Cambridge MA.

te Velde, J.R. 2006. *Deriving Coordinate Symmetries.* Benjamins, Amsterdam.

Yatabe, S. 2001. The syntax and semantics of left-node raising in Japanese. *Procs. of the 7th Int. HPSG Conf.,* Berkeley. CSLI Publications, Stanford CA.