A comparison of clausal coordinate ellipsis in Estonian and German: Remarkably similar elision rules allow a language-independent ellipsis-generation module

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

Mare Koit & Haldur Õim
Research Group of Computational Linguistics
University of Tartu
Tartu, Estonia
mare.koit@ut.ee & haldur.oim@ut.ee

Abstract

We compare the phenomena of clausal coordinate ellipsis in Estonian, a Finno-Ugric language, and German, an Indo-European language. The rules underlying these phenomena appear to be remarkably similar. Thus, the software module ELLEIPO, which was originally developed to generate clausal coordinate ellipsis in German and Dutch, works for Estonian as well. In order to extend ELLEIPO's coverage to Estonian, we only had to adapt the lexicon and some syntax rules unrelated to coordination. We describe the language-independent rules for coordinate ellipsis that ELLEIPO applies to non-elliptical syntactic structures in both target languages.

1 Introduction

In written German newspaper text, clausal coordination occurs in about 14% of the sentences, and coordinate ellipsis (e.g. (1)) in about 7% (see a corpus study by Harbusch and Kempen, 2007). Studies of ellipsis in Estonian are hardly available (cf. Erelt, 2003).

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

In order to deal with these relatively frequent phenomena, we develop an Estonian coordinate-ellipsis generator based on ELLEIPO, the software module written in JAVA that generates clausal coordinate ellipsis in German and Dutch (Harbusch and Kempen, 2006; 2009). Given the fact that the two target languages belong to two rather different language families (German is an Indo-European, Estonian a Finno-Ugric language) we expected the two target languages to differ considerably with respect to the rules for generating coordinate elisions; however, this expectation was falsified. As we will detail below, a pairwise comparison of a heterogeneous set of elliptical constructions in the target languages reveals that the German rules we had implemented in ELLEIPO also generate the Estonian structures. We only needed to adapt the lexicon and some syntax rules unrelated to coordination. The core algorithm worked language-independently for both languages.

The paper is organized as follows. In section 2, we first define the four main groups of clausal coordinate ellipsis phenomena, and show that the elisions in the two target languages obey basically the same rules. This implies that the Estonian version of the software system ELLEIPO can use the same core algorithm as the German and Dutch version. In section 3, we discuss other linguistic theories for clausal coordinate ellipsis, especially focussing on implementations for generation. In final section 4, we draw some conclusions and address options for future work.

2 Clause-level coordinate ellipsis in Estonian and German

In the literature, one often distinguishes four major types of clause-level coordinate ellipsis (which can become combined; cf. example (1)).[1]

- GAPPING, with three special variants called LONG DISTANCE GAPPING (LDG), SUB-GAPPING, and STRIPPING,
- FORWARD CONJUNCTION REDUCTION (FCR),
- BACKWARD CONJUNCTION REDUCTION (BCR;

[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., *John likes skating and Peter likes skiing* becoming *John and Peter like skating and skiing, respectively*). Presumably, such conversions involve a logical rather than syntactic mechanism.

Proceedings of the EACL 2009 Demonstrations Session, pages 25–28,
Athens, Greece, 3 April 2009.

also called *Right Node Raising*), and

- SUBJECT GAP IN CLAUSES WITH FINITE/ FRONTED VERBS (SGF).

They are illustrated in the English sentences (2) through (8). The subscripts denote the elliptical mechanism at work: *g* stands for Gapping, Subgapping, and Stripping, respectively; $g(g)^+$ is recursively added for LDG; *f* = FCR; *s* = SGF; *b* = BCR.

(2) GAPPING*: Jüri* ***lives*** *in Tallinn and his children* ~~***live***~~$_g$ *in Tartu*

(3) LDG: *My wife* ***wants to buy*** *a car and my son* ~~***wants***~~$_g$ ~~***[to buy]***~~$_{gg}$ *a motorcycle*

(4) SUBGAPPING: *The driver* ***was*** *killed and the passengers* ~~***were***~~$_g$ *severely wounded*

(5) STRIPPING*: My sister* ***lives in Narva*** *and my brother* ~~***[lives in Narva]***~~$_g$ ***too***

(6) FCR: *Pärnu is the city [S* ***where*** *Ainar lives and* ~~***where***~~$_f$ *Peeter works]*

(7) BCR: *Riina arrived before three* ~~***[o'clock]***~~$_b$ *and Terje left after six* ***o'clock***

(8) SGF: *Into the wood went* ***the hunter*** *and* ~~***[the hunter]***~~$_s$ *shot a hare*

In the theoretical framework by Kempen (2009) and its implementation for German and Dutch in ELLEIPO, the elision process is guided by constraints on *lemma-* and *wordform-identity* constraints and, to some extent, linear order.[2]

ELLEIPO's functioning is based on the assumption that coordinate ellipsis does not result from the application of declarative grammar rules for clause formation but from a procedural component that interacts with the sentence generator and may block the overt expression of certain constituents. Thus, the rules apply to assembled non-elliptical (unreduced) tree structures in the final stage of generation. Due to this feature, ELLEIPO can be combined, at least in principle, with various lexicalized-grammar formalisms. However, this advantage does not come entirely for free: The module needs a formalism-dependent interface that converts generator output to a *canonical form* consisting of "flat" syntactic trees where all major clause constituents are represented at the same hierarchical level (see Harbusch and Kempen 2006; 2007).

In the following, we introduce ELLEIPO's elision rules only in an informal manner (for the pseudocode of the algorithm, see Harbusch and Kempen, 2006; 2009). The rules described in the following can be applied in any order to unreduced syntactic structures in canonical form. In case of a successful rule application, the elidable constituents (and its non-elided counterpart in the other conjunct) is adorned with a subscript indicating the ellipsis type (as illustrated in (2) through (8)). ELLEIPO's final step executes all possible elliptical combinations (e.g., for example (1), it also realizes a version with Subgapping and LDG, respectively, i.e.: *Monopole* ***sollen*** *geknackt* ***werden*** *und Märkte* ~~***sollen***~~$_g$ *getrennt* ~~***werden***~~$_{gg}$**)**.

In Gapping (see examples (9) and (10)), lemma-identical verbs can be elided from the second conjunct, if and only if a contrast is expressed, i.e. each remaining constituent in this conjunct has a counterpart with the same grammatical function in the first conjunct (cf. (11)).[3]

(9) *Mari* ***loeb*** *artikleid ja tema pojad* $\__g$ *pakse raamatuid*
Mari ***liest*** *Artikel und ihre Söhne* $\__g$ *dicke Bücher*
Mari reads articles and her sons thick books

(10) *Jüri* ***elab*** *Tartus ja Tallinnas* $\__g$ *tema pojad*
Jüri ***lebt*** *in Tartu und in Tallinn* $\__g$ *seine Söhne*
Jüri lives in Tartu and in Tallinn his sons

(11) **Mari* ***ostab*** *pirne ja Jüri* $\__g$ *turul*
Mari* ***kauft* *Birnen und Jüri* $\__g$ *auf dem Markt*
Mari buys pears and Jüri on the market

In *Long-Distance Gapping (LDG)*, the *remnants, i.e.* the non-elided constituents in the posterior conjunct, include constituents whose anterior counterparts belong to different clauses. *My wife* in (12) (translation of (3)) belongs to the main clause whereas *a car* is part of the infinitival complement clause. Notice that LDG does not require adjacency of the elided verbs (cf. the German example in (12)).

(12) *Minu naine* ***soovib osta*** *autot ja minu poeg* ~~***soovib***~~$_g$ ~~***osta***~~$_{gg}$ *mootorratast*
Meine Frau ***will*** *ein Auto* ***kaufen*** *und mein Sohn* ~~***will***~~$_g$ *ein Motorrad* ~~***kaufen***~~$_{gg}$

In *Subgapping*, the posterior conjunct includes a remnant in the form of a non-finite complement

[2] Coordinate structures consist of two or more *conjuncts* connected by a coordinating conjunction (in our examples: *and*). Rules of coordinate ellipsis license elision of some consituent in one conjunct under "identity" with a constituent in another conjunct. We distinguish between *lemma identity,* where only the word-stems of the constituents have to be identical, and *wordform identity,* which requires not only identity of the stems but also of their morphological features. Gapping only requires lemma identity (cf. examples (2) and (4)). In FCR, word-form identity is checked, i.e. the identical word string referring to the same referent (cf. ****The boy*** *loves dogs and* ~~***[the boys]***~~$_f$ *hate cats*).

[3] For lack of space, here we cannot go into aspects of word-order variation (both Estonian and German are languages with relatively free word order). For the same reason, we only discuss examples with two conjuncts (although, ELLEIPO analyses *n*-ary coordinations as well), and cannot pay attention to coordinate structures that include negation.

clause ("VP"; *severely wounded* in (13); translation of (4)).

(13) *Juht* ***sai*** *surma ja reisijad* **_g** *tõsiselt vigastada*
Der Fahrer ***wurde*** *getötet und die Passagiere*
_g *ernsthaft verletzt*

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 (14), the translation of (5)).

(14) *Mu õde* ***elab Narvas*** *ja mu vend* **_g** *samuti/ka.*
Meine Schwester ***lebt in Narva*** *und mein Bruder*
_g *ebenso/ auch*

In *Forward-Conjunction Reduction (FCR)*, a left-peripheral string of major constituents in the right conjunct is elided under wordform-identity with its counterpart in the right conjunct. In FCR example (15), the left-peripheral string comprising complementizer, subject and direct object are elided from the right-hand conjunct. If modifiers that are neither lemma- nor wordform-identical, are placed in between subject and object—as in (16)—, then elision of the object is blocked. (Actually, example (16) is not ill-formed but its right-hand conjunct cannot be interpreted as *cleaning the bike*.) In main-clause variant (17), elision of the direct object is blocked for similar reasons.

(15) *...* ***et Jan oma jalgratta*** *asjatundlikult parandas*
... ***dass Jan sein Fahrrad*** *fachkundig reparierte*
... that Jan his bike expertly repaired
ja ***[et Jan oma jalgratta]***$_f$ *hoolikalt puhastas*
und ***[dass Jan sein Fahrrad]***$_f$ *eifrig putzte*
and that Jan his bike diligently cleaned

(16) **...* ***et Jan*** *asjatundlikult* ***oma jalgratta*** *parandas*
... ***dass Jan*** *fachkundig* ***sein Fahrrad*** *reparierte*
ja ***[et Jan]***$_f$ *hoolikalt* ***[oma jalgratta]***$_f$ *puhastas*
und ***[dass Jan]***$_f$ *eifrig* ***[sein Fahrrad]***$_f$ *putzte*

(17) * ***Jan*** *parandas* ***oma jalgratta*** *asjatundlikult*
* ***Jan*** *reparierte* ***sein Fahrrad*** *fachkundig*
ja ***Jan***$_f$ *puhastas* ***[oma jalgratta]***$_f$ *hoolikalt*
und ***Jan***$_f$ *putzte* ***[sein Fahrrad]***$_f$ *eifrig*

Backward-Conjunction Reduction (BCR) licenses elision of a right-peripheral string in the left-hand conjunct under lemma-identity[4] with its counterpart in the right conjunct. However, unlike FCR's mirror image, BCR may cut into major constituents of the clause. In BCR example (18), the direct object can be elided in the first conjunct whereas in word-order variant (19), the verb blocks this elision. Example (20) illustrates that BCR, unlike the three other ellipsis types, may cut into major clausal constituents and only checks lemma-identity. Varying the objects to *'new bike'/'old bikes'*, and the second subject *'Peter'* to *'his brothers'* does not rule out ellipsis as long as peripheral access is guaranteed.

(18) *Jan parandas* ***[oma jalgratta]***$_b$
Jan reparierte ***[sein Fahrrad]***$_b$
Jan repaired his bike
ja Peeter puhastas ***oma jalgratta***
und Peter putzte ***sein Fahrrad***
and Peter cleaned his bike

(19) **... et Jan* ***[oma jalgratta]***$_b$ *parandas*
* *... dass Jan* ***[sein Fahrrad]***$_b$ *reparierte*
ja et Peeter ***oma jalgratta*** *puhastas*
und dass Peter ***sein Fahrrad*** *putzte*

(20) *Jan parandas* ***oma uue jalgratta***$_b$
Jan reparierte ***sein neues Fahrrad***$_b$
ja tema vennad puhastasid ***oma vanad jalgrattad***
und seine Brüder putzten ***ihre alten Fahrräder***

Examples (21)-(23) embody word-order variants within two simple coordinated clauses. The (il)licit elision patterns verify that in BCR the ellipsis should be right-peripheral in the left-hand conjunct, whereas in FCR the ellipsis is located left-peripherally in the right-hand conjunct.

(21) *Mari loeb* **_**$_b$ *ja Jüri kirjutab* ***raamatuid***
Mari liest **_**$_b$ *und Jüri schreibt* ***Bücher***
Mari reads and Jüri writes books

(22) ***_**$_b$ *Loeb Mari ja* ***raamatuid*** *kirjutab Jüri*
_**$_b$ *Liest Mari und* ***Bücher *schreibt Jüri*
reads Mari and books writes Jüri

(23) ***Raamatuid*** *loeb Mari ja* **_**$_f$ *kirjutab Jüri*
Bücher *liest Mari und* **_**$_f$ *schreibt Jüri*
Books reads Mari and writes Jüri

SGF (Subject Gap in clauses with Finite/Fronted verb) licenses elision of the subject of the right conjunct if in the left conjunct the subject follows the verb; however, the first constituent of the unreduced right-hand clausal conjunct must meet certain special requirements. In particular, it should be the subject of this clause (as in (24), translation of (8)) or a modifier (25), but not an argument other than the subject, e.g. neither complement nor (in)direct object (26). Additionally, if FCR is also possible, it should actually be realized in order to license SGF (for additional discussion of these restrictions, see Harbusch and Kempen, 2009).

(24) ***Metsa*** *läks* ***jahimees*** *ja* **_**$_s$ *tappis jänese*
In den Wald *ging* ***der Jäger*** *und* **_**$_s$ *schoss einen Hasen.*

(25) ***Miks/Eile*** *oled* ***sa*** *läinud ja*
Warum *bist* ***du*** *gegangen und*
Why have you left and
_$_f$ *ei ole* **_**$_s$ *midagi* *öelnud?*
_$_f$ *hast* **_**$_s$ *mich nicht* *gewarnt?*
have not me (Est.)/have me not (Ger.) warned
'Why did you leave but didn't you warn me?'

[4] ELLEIPO also checks case-identity to rule out *?Hilf* **_**$_{b[DAT]}$ *und reanimier [den Mann]*$_{ACC}$ 'Help and reanimate the man'

(26) ***Seda veini** *ei joo* **ma**
***Diesen Wein** *trinke* **ich** *nicht*
This wine drink not I (Est.)/drink I not (Ger.)
enam ja ~~***[selle veini]***~~$_f$ *kallan* ~~***ma***~~$_s$ *ära*
mehr und ~~***[diesen Wein]***~~$_f$ *gieße* ~~***ich***~~$_s$ *weg*
anymore and this wine throw I away
'I don't drink this wine and throw it away'

Given the similarities between the rules that appear to control clausal coordinate ellipsis in German and Estonian, it is not surprising that the German/Dutch version of ELLEIPO could be tailored to Estonian easily. ELLEIPO's language-independent core algorithm generates Estonian ellipsis as well, as shown by the demonstrator. For the sake of completeness, we should add here that we have not been able to find types of clausal coordinate ellipsis in Estonian that go beyond the above four types; hence, as far as we can tell, Estonian does not require additional rules over and above those we needed for German and Dutch.

3 State of the art in ellipsis generation

All major grammar formalisms provide rules for clausal coordinate ellipsis—rules that tend to be intertwined with rules for nonelliptical coordination (e.g. Sarkar and Joshi (1996) for Tree Adjoining Grammar; Steedman (2000) for Combinatory Categorial Grammar; Frank (2002) for Functional Grammar; Crysman (2003) and Beavers and Sag (2004) for HPSG; and te Velde (2006) for the Minimalist Program). This also applies to many NLG systems (cf. Reiter and Dale, 2000). Generators that do include an autonomous component for coordinate ellipsis—that is, a component that takes unreduced coordinations expressed in the system's grammar formalism as input and return elliptical versions as output (Shaw, 1998; Dalianis, 1999; Hielkema, 2005)—use incomplete rule sets, thus risking over- or undergeneration, and incorrect or unnatural output.

4 Conclusion

Finally, we do not expect that the four types of clausal coordinate ellipsis presented here are "universal" in the sense that all natural languages exhibit all four of them and no language has additional types (see Harbusch and Kempen 2009 for some discussion based on language-typological work by Haspelmath, 2007). However, the experience described in this paper makes us confident that the "modular" approach taken in the ELLEIPO project will prove efficient when it comes to writing coordinate ellipsis rules for other languages—especially for languages belonging other language families.

References

John Beavers and Ivan A. Sag. 2004. Coordinate Ellipsis and Apparent Non-Constituent Coordination. In: *Procs. of 11th Int. HPSG Conf.*, Leuven, 48-69.

Hercules Dalianis. 1999. Aggregation in natural language generation. *Computational Intelligence, 15:* 384-414.

Berthold Crysmann. 2003. An asymmetric theory of peripheral sharing in HPSG. In: *Procs. of 8th Conf. on Formal Grammar*, Vienna.

Mati Erelt (Ed.). 2003. *Estonian Language*. Estonian Academy Publishers, Tallinn.

Anette Frank. 2002. A (discourse) functional analysis of asymmetric coordination. In*: Procs. of the LFG02 Conf.*, Athens, pp. 174-196.

Karin Harbusch and Gerard Kempen. 2006. ELLEIPO: A module that computes coordinate ellipsis for language generators that don't. In: *Procs. of 11th EACL,* Trento, pp. 115-118.

Karin Harbusch and Gerard Kempen. 2007. Clausal coordinate ellipsis in German. In: *Procs. of 16th NODALIDA*, Tartu, pp. 81-88.

Karin Harbusch and Gerard Kempen. 2009. Generating clausal coordinate ellipsis multilingually. In: *Procs. of 12th ENLG*, Athens.

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

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

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

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems.* Cambridge University Press, Cambridge, UK.

Anoop Sarkar and Aravind Joshi. 1996. Coordination in Tree Adjoining Grammars: Formalization and implementation. In: *Procs. of 16th COLING,* Copenhagen, pp. 610–615.

James Shaw. 1998. Segregatory coordination and ellipsis in text generation. In: *Procs. of 17th COLING,* Montreal, pp. 1220-1226.

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

John R. te Velde. 2006. *Deriving Coordinate Symmetries: A Phase-Based Approach Integrating Select, Merge, Copy and Match*. John Benjamins, Amsterdam.