Rule-Based Detection of Clausal Coordinate Ellipsis

Kristiina Muhonen & Tanja Purtonen

University of Helsinki
Department of Modern Languages
FIN-CLARIN
{kristiina.muhonen, tanja.purtonen}@helsinki.fi

Abstract

With our experiment, we show how we can detect and annotate clausal coordinate ellipsis with Constraint Grammar rules. We focus on such an elliptical structure in which there are two coordinated clauses, and the latter one lacks a verb. For example, the sentence *This belongs to me and that to you* demonstrates the ellipsis in question, namely gapping. The Constraint Grammar rules are made for a Finnish parsebank, FinnTreeBank. The FinnTreeBank project is building a parsebank in the dependency syntactic framework in which verbs are central since other sentence elements depend on them. Without correct detection of omitted verbs, the syntactic analysis of the whole sentence fails. In the experiment, we detect gapping based on morphology and linear order of the words without using syntactic or semantic information. The test corpus, Finnish Wikipedia, is morphologically analyzed but not disambiguated. Even with an ambiguous morphological analysis, the results show that 89,9% of the detected sentences are elliptical, making the rules accurate enough to be used in the creation of FinnTreeBank. Once we have a morphologically disambiguated corpus, we can write more accurate rules and expect better results.

Keywords: Ellipsis, Treebanks, Constraint Grammar

1. Introduction

Ellipsis in coordinated clauses is a widely known and discussed linguistic issue. In syntactic parsing and generation, it raises at least two kinds of problems: first, it can be hard to detect automatically, and second, it can be difficult to model in a treebank or a parsebank. In this article, we focus particularly on the first problem, detecting the phenomenon automatically, but we also define an annotation scheme for the coordinated elliptical clause type in focus: GAPPING. In gapping, two clauses are coordinated so that the posterior conjunct lacks a verb, like in the sentence: *Some are positive and some negative.*

We approach the problem with a rule-based method, Constraint Grammar (CG)[1] (Karlsson et al., 1995). We show how gapping can be detected and consequently annotated with a brief and efficient grammar.

The CG grammar created for the experiment is used in building FinnTreeBank, a dependency treebank/parsebank for Finnish (Voutilainen et al., 2011). FinnTreeBank is part of the Finnish CLARIN infrastructure, FIN-CLARIN[2], and provides language resources for researchers by creating a manually annotated treebank, an automatically created parsebank, and a dependency parser for unrestricted text. The first manually annotated version of the treebank is already published, and currently the FinnTreeBank project is creating a parsebank using Constraint Grammar rules for morphological disambiguation and syntactic parsing.

In the experiment described in this paper, we detect clausal coordinate ellipsis from the Finnish Wikipedia using CG rules. After detection, we manually analyze the output of the grammar to estimate the accuracy of automatic detection of elliptical structures. If the rules prove to be accurate enough, we can build on them when creating the dependency syntactic annotation for the parsebank. The results show how precisely we can detect elliptical coordinated clauses based on morphological information in a rule-based way.

[1]The latest CG compiler, VISL CG-3 (Didriksen, 2011), is available for download here:
http://beta.visl.sdu.dk/cg3.html

[2]http://www.ling.helsinki.fi/finclarin/

2. Modeling Clausal Coordinate Ellipsis

As Hakulinen and Karlsson (1988) report, in Finnish, there are at least three main types of ellipsis: ellipsis of the main word, conjunction reduction, and gapping. In this paper, we focus on gapping. It differs from the other ellipsis types so that in gapping, the verb of the posterior conjunct is omitted. FinnTreeBank is building the parsebank in the dependency syntactic framework in which verbs are central. Hence, it is crucial to detect the omitted verbs already at the syntactic level to ensure correct analyses of the sentence.

When building a parsebank, ellipsis that does not involve the verb, e.g. an omitted object, can be left invisible on the syntactic level. Nominal ellipsis does not necessarily cause problems in annotating the rest of the sentence correctly. However, undetected omitted verbs can lead to incorrect analyses of the whole sentence. For example, a subject and an object are dependents of the verb, so their dependency relations in the elliptical clause cannot be analyzed correctly without detecting the omitted verb.

2.1. Gapping

The elliptical structure we focus on in our experiment is a type of clausal coordinate ellipsis: gapping (Hakulinen and Karlsson, 1988, p. 324). Harbusch and Kempen (2007) give an overview on elliptical coordination in English and German, the phenomenon being very similar to ellipsis in Finnish. For coordinate ellipsis in other languages, see e.g Haspelmath (2004).

In gapping, the posterior conjunct of a coordinated sentence lacks a verb, and the main verb is borrowed from the anterior conjunct. The whole finite verb (with its auxiliaries) is missing, distinguishing it from such verbal ellipsis in which the auxiliary is not omitted, e.g. *She has been to Sweden and he has not.*
Example (1) from the Finnish Wikipedia demonstrates gapping in Finnish.

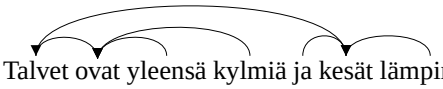
(1) Päälaki on tasainen ja silmät suuret.
vertex-NOM is flat-NOM and eyes-NOM big-NOM
The vertex is flat and the eyes big.

Example (1) portrays the elliptical coordination we capture with the CG rules. In the posterior elliptical clause *silmät suuret (eyes big)*, the verb *on (is)* is omitted and borrowed from the main clause. Otherwise, the clause contains the same sentence elements as the main clause: the subject *silmät (eyes)* and the adjectival predicative *suuret (big)*, both in the nominative case (NOM). As can be seen in Example (1), the verb does not have to occur in the same number in the two conjuncts: the actual verb in the anterior conjunct *on (is)* is in singular and the omitted verb of the posterior conjunct *ovat (are)* is in plural.

2.2. FinnTreeBank's Annotation Scheme

There is no straightforward way of parsing elliptical clauses in the dependency syntactic framework in which e.g. an object and a subject are always dependents of the verb. Generally, there are two main approaches to portraying elliptical elements in treebanks: adding the unrealized, omitted word, and then annotating the completed sentence (see e.g. Stanford scheme (de Marneffe et al., 2006) and (de Marneffe and Manning, 2008)), or annotating only realized words on the syntactic level (see e.g. Prague dependency treebank (Hajič, 1998)).
The FinnTreeBank project is building large-scale annotated corpora of authentic language. Therefore, we do not adopt the approach in which the sentences are modified and e.g. the "missing" verbs added. In other words, the annotation scheme of FinnTreeBank is based on surface syntax.
Example (2) demonstrates FinnTreeBank's annotation scheme for gapping.

(2) Talvet ovat yleensä kylmiä ja
winters-NOM are generally cold-PAR and
kesät lämpimiä.
summers-NOM warm-PAR
Winters are generally cold and summers warm.



We coordinate the first sentence element of the posterior conjunct with the morphologically equivalent sentence element in the main clause. Usually, and always in the experiment reported here, it means that we coordinate the subjects.
The subject of the elliptical posterior conjunct *kesät (summers)* is seen as a direct dependent of the subject of the main clause *talvet (winters)*, and its function is COORDINATED ELLIPTICAL SUBJECT. The other sentence elements of the elliptical clause are directly linked to the subject of the elliptical clause: the partitive (PAR) predicative *lämpimiä (warm)* is a dependent of *kesät (summers)*.
Gapping can also occur in sentences with an elliptical subject (Hakulinen and Karlsson, 1988, p. 325). Though the main focus of this paper is on gapping, we portray the annotation scheme for the co-occurrence of gapping and an elliptical subject in Example (3).

(3) Hän lukee aamulla lehteä ja illalla
she reads morning-ADE paper-PAR and evening-ADE
kirjaa.
book-PAR
She reads the paper in the morning and a book in the evening.

The principles we follow for constructing the annotation scheme for gapping can also be applied in Example (3). The posterior conjunct is elliptical in two ways: it lacks both the verb *lukee (reads)* and the subject *hän (she)*. When gapping co-occurs with an elliptical subject, it is impossible to coordinate the subjects. In such cases, the first morphologically similar counterpart of the posterior conjunct *illalla (evening)* is coordinated with its counterpart in the main clause *aamulla (morning)*, both in the adessive case (ADE).
We ended up with the annotation solution described in Example (3) after consulting the future users of FinnTreeBank on the most intuitive annotation scheme for elliptical comparative clauses (Muhonen and Purtonen, 2011). The results of the user query suggest that the dependency is seen between the first equivalent words most frequently.

3. Rule-Based Detection of Gapping

We will now move forward from the linguistic definition of gapping towards the rule based implementation of the phenomenon. We assume that since gapping can be defined linguistically, it can also be parsed using Constraint Grammar.
In Finnish, many of the elliptical contexts, including gapping, can be defined with the help of case markers, but e.g. in English, the same can be done with prepositions, like in the following sentence:

(4) This belongs **to** me and that **to** you.

The rule-based approach enables detecting elliptical coordination which can be difficult to parse correctly with statistical methods. For example, the Stanford parser[3] (de Marneffe et al., 2006) parses Example (4) so that the word chain *"me and that"* forms an NP.

[3] http://nlp.stanford.edu:8080/parser/

3.1. Linguistic Cues

Before we can write the Constraint Grammar for capturing gapping, we need to define the linguistic environment in which gapping occurs. In gapping, the elliptical clause contains at least two sentence elements that have counterparts in the main clause. We can thus detect the elliptical clauses based on the similarity of these counterparts.

Since we only have a morphologically analyzed corpus available that lacks any syntactic or semantic analysis, we have to base the detection of gapping solely on morphology. This sets restrictions on how expressive the CG rules can be and forces us to simplify the linguistic phenomena. Hence, the elliptical structures that we detect with the CG rules fit the following template: the first word is a subject of the main clause in the nominative case. The second word the rules find is an object, adverbial, or a predicative. This word has to be inflected in the same grammatical case as its counterpart in the main clause. Example (5) demonstrates a simplified example from the Wikipedia.

(5) Korkeus on 0,65 m ja leveys 0,60 m.
height-NOM is 0.65 m and width-NOM 0.60 m
The height is 0.65 m and the width 0.60 m.

In Example (5), the morphological cues for finding gapping are so clear that we can mark such structures with CG rules. We have to fix the linear order of words in the structure and set restrictions on what can occur between the two subjects in the nominative case.

Since we do not have a morphologically disambiguated corpus, where e.g. all adverbials would be marked, we cannot be sure of the dependency functions of the words. Finnish is a free constituent order language so the functions cannot be solved based on word order. To avoid erroneous analyses caused by this, we do not allow for anything to occur between the subject of the posterior conjunct and the conjunction or comma. That is, the word *leveys (width)* needs to directly follow the conjunction *ja (and)*.

In elliptical coordinated structures, the posterior conjoined clause usually contains words semantically related to their equivalents in the main clause. In Example (5), these semantically equivalent words are *korkeus (height)* and *leveys (width)*, and *0,65 m* and *0,60 m*. The existence of such counterparts lead us to assume that it would be easier to detect gapping if we could use semantic information in addition to morphology. However, at present, the treebank/parsebank does not contain any semantic or tectogrammatical level, and we aim at a precise and informative analysis already at the syntactic level. Hence, in this paper we test how precisely we can detect elliptical coordinated clauses only based on morphological information and the linear order of words.

3.2. CG Experiment

To detect gapping, we created a short Constraint Grammar. The rules add a tag "`ELL_SUBJ`" to the subject of the posterior conjoined clause. In Example (2), the rules add the tag to the word *kesät (summers)* indicating that the word is the subject of the elliptical clause. Analogously, in Example (5), the rules tag *leveys (width)* as the subject of the posterior elliptical conjunct. The purpose of the experiment was not to capture each ellipsis type or to examine the frequency of the phenomena, but to demonstrate how CG rules can be used for detecting and annotating gapping.

The grammar contains two rules. However, the context conditions of CG rules are practically arbitrarily complex, so that the number of rules is not a good indicator of grammar coverage and complexity. Simply put, our rules capture gapping in the posterior conjunct described in Section 2: first a subject in the nominative case, then an object, adverbial, or a predicative in the same case as its counterpart in the main clause.

We tested our rules on the body text of the Finnish Wikipedia. The test corpus was a short extract (2%) from the Finnish Wikipedia. The rules were optimized to cover the phenomenon in the test corpus after several test runs

After the rules were optimized for the test corpus, they were applied to the whole Finnish Wikipedia. The corpus was preprocessed and morphologically analyzed using OMorFi (Pirinen, 2011), but not disambiguated. This means that words can have several morphological analyses of which only one is correct. Since the CG rules are based on grammatical cases, this causes problems. CG offers a special operator, the `C flag`, for restricting the rules to work on only such words that have a "safe" reading. Bick (2009) defines the safe flag as follows: *"A C (careful) condition attached to the position number means that the context condition has to be a safe (i.e. the only) reading of the cohort in question."* We use this option e.g. when finding nominative subjects: the context condition (1C N Nom) denotes an unambiguous noun in the nominative case to the right, that is, the right adjacent word. E.g. a word with both a noun and a verb reading in this position violates the context condition.

The context conditions in the CG grammar can be defined in an arbitrarily complex way so that the rules return structures that match very specific criteria. At this stage of development, the rules are defined so that they only match such occurrences of gapping that can be detected with morphological information only. Hence, the rules cover such gapping that is explicitly defined with the Constraint Grammar.

Since we require the semantically equivalent words, e.g. the adverbials of the two conjuncts, to be in the same grammatical case, we cannot capture all valid occurrences of gapping. Such a sentence is portrayed below in Example (6).

(6) Minä menen Espooseen ja sinä Vantaalle.
I-NOM go Espoo-ILL and you-NOM Vantaa-ALL
I go to Espoo and you to Vantaa.

In Finnish, the grammatical case of locative expressions containing proper nouns like city names differ from each other. Example (6) shows how Espoo and Vantaa are inflected in different locative cases when indicating direction of movement. Espoo inflects in the illative (ILL) case ("into (the inside of))", while Vantaa inflects in the allative (ALL) case ("onto"). The reason for this is that the inflection patterns have become established in the language and follow no particular pattern. In Example (6), if the city names

Espoo and Vantaa would be analyzed as adverbials of location, we could base the CG rules on this information as well and broaden the coverage of our rules.

4. Results and Further Remarks

We will now move on to discussing the results of the CG experiment. After calibrating the rules by running them on the test corpus, we ran the rules on the whole Finnish Wikipedia. Since gapping as a phenomenon is rather rare, we emphasize qualitative evaluation of the results.

4.1. Success Rate

The CG rules captured gapping 1 333 times from the Finnish Wikipedia. We evaluated the sentences captured by the rules by hand to see whether all 1 333 sentences actually are elliptical. Manual evaluation enables accurate classification of errors and allows us to assess the rules better.
The results of the experiment show that already a brief CG grammar succeeds in finding coordinate ellipsis. The results together with an error classification are shown in Table 1.

HITS	N	%
Correct	1 197	89,9%
Title Error	99	7,4%
Other Error	37	2,8%
Total	1 333	100%

Table 1: Results

The CG rules succeed in revealing gapping in 1 197 cases out of the 1 333 retrieved sentences. This means that the success rate of the rules is 89,9%.

4.2. Error Analysis

Errors occur most frequently in sentences with titles and compounds, e.g. "Nobelist Curie" or "Playstation 3". Such sentences make up 73% of the errors. The frequency of these title errors can be explained by the fact that in Finnish, titles like "Nobelist" are not capitalized, making recognizing them more challenging. The rest, 27%, are miscellaneous mistakes where the rules cannot distinguish gapping correctly. Fixing these mistakes requires a thorough scrutiny of the context conditions of the rules.
Preliminary tests proved that proper names pose a similar problem to the title errors ("Nobelist Curie"). Sentences like *"He meets me and John Lee."* are structurally ambiguous: Is the posterior conjunct *"John Lee"* a coordinated elliptical clause, or does *"me and John Lee"* form an object NP? Initially, we ran the rules on the test corpus and saw that most of the mistakes (∼80%) occurred in sentences with proper names in the elliptical sentence. Because at this stage we do not want to focus on the ambiguity problem caused by proper names, we calibrated the rules so that sentences with them were disregarded. Should we have a finer-grained classification of proper names at our disposal, we could take them into account as well and capture more results with our rules.
27 % of the errors are classified "other errors". These errors are mostly caused by unspecific context conditions of the rules. We used a relatively small test corpus (2% of the Finnish Wikipedia) to optimize the rules. We corrected all other errors but title errors in the test phase by writing more accurate context conditions. However, the test corpus did not contain all structures that can be misinterpreted as gapping.
Out of the 37 "other errors", 14 were caused by the rules allowing any verb to occur after the equivalent words in the posterior conjunct. We did not limit the occurrence of verbs with stricter conditions because without morphological disambiguation some nouns and adverbs can have a verb reading as well. However, we did not mind some non-elliptical coordinated clauses beginning with the same sentence elements as elliptical coordinated clauses. This causes the following kind of erroneous gapping discoveries:

(7) Kilpailu koostui 16 lajista, osa
competition-NOM comprise 16 sport-ELA part-NOM
lajeista suoritettiin kahdesti.
sport-ELA carried out twice
The competition comprises 16 sports, some sports were carried out twice.

In Example (7), the latter clause *osa lajeista suoritettiin kahdesti* contains a verb and is not elliptical. Therefore it should not be analyzed as gapping. The error is caused by the nouns in both conjuncts inflecting in the same cases: a noun in the nominative case followed by a noun in the elative (ELA) case. Moreover, since we did not restrict the occurrence of a verb after the equivalent words in the posterior conjunct, the structure is falsely recognized as gapping.
Using the `C flag` of VISL CG-3, we can make the context conditions of the posterior conjunct stricter and exclude words with only verb readings. With this improvement, we can rule out 13 mistakes, fixing 10% of the errors in the corpus.
Other errors are more arbitrary, but we can make some general remarks on them as well. Many of the errors were uninteresting from the syntax's point of view. For example, the morphological analyzer does not recognize the case inflection of abbreviations correctly, but analyzes every abbreviation as a nominative. These wrong case markers result in incorrect gapping discoveries.
In addition to the errors caused by incorrect morphology or incomplete context conditions of the rules, there are some structures that are more complex to solve. For example, the following sentence is erroneously marked as gapping, and it is difficult to distinguish from the elliptical structure automatically, e.g.:

(8) Tuli tuhosi osan hyteistä ja puolet
fire-NOM destroyed part cabins-ELA and half-NOM
ruokasalista.
dining hall-ELA
The fire destroyed some cabins and half of the dining hall.

Example (8) is a non-elliptical sentence with coordination. The two noun phrases *osan hyteistä (some cabins)* and *puolet ruokasalista (half of the dining hall)* are coordinated with each other. The rules analyze this coordination erroneously as gapping. The error is caused by the two-word noun phrase *puolet ruokasalista*. In Finnish, two nouns in different grammatical cases do not usually form an NP. However, case government overrides this tendency. The noun that follows the word *puolet (half)* must be in the elative case. These kinds of NPs with case government must be identified before completing gapping detection so that errors like in Example 8 can be fixed.

5. Conclusion

In this experiment, we detected elliptical structures from an morphologically undisambiguated corpus without any semantic or syntactic information. The elliptical structure we focused on is gapping, and we used a rule-based method, Constraint Grammar. The detection is based on the grammatical cases and the linear order of the words.

In 89,9% of the sentences captured by the Constraint Grammar rules, the structure is analyzed correctly as gapping. Most errors occur in sentences with titles and compounds. If the corpus we detect gapping from would be morphologically disambiguated and the gapping rules would be developed alongside with other syntactic rules (e.g. recognition of titles), we could expect better results.

The rules are written to work on a corpus that has only limited linguistic annotation. Nonetheless, the results encourage us to build on the preliminary rules written for this experiment when modeling clausal coordinate ellipsis in FinnTreeBank.

Acknowledgements

The ongoing project has been funded via CLARIN, FIN-CLARIN, FIN-CLARIN-CONTENT and META-NORD by EU, University of Helsinki, and the Academy of Finland. We would like to thank the three anonymous reviewers for their constructive comments.

6. References

Eckhard Bick, 2009. *Basic Constraint Grammar Tutorial for CG-3 (Vislcg3)*. Visl, Syddansk Universitet. http://beta.visl.sdu.dk/cg3_howto.pdf.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual. Technical report, Stanford University.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC 2006*.

Tino Didriksen. 2011. Constraint Grammar Manual: 3rd version of the CG formalism variant. GrammarSoft ApS. http://beta.visl.sdu.dk/cg3/vislcg3.pdf.

Jan Hajič. 1998. Building a syntactically annotated corpus: The Prague dependency treebank. *Issues of valency and meaning*, pages 106–132.

Auli Hakulinen and Fred Karlsson. 1988. *Nykysuomen lauseoppia*. SKS, Helsinki, 2 edition. ISBN 951-717-543-4.

Karin Harbusch and Gerard Kempen. 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)*.

Martin Haspelmath, editor. 2004. *Coordinating Constructions*. Typological Studies in Language 58. John Benjamins, Philadelphia.

Fred Karlsson, Atro Voutilainen, Juha Heikkilä, and Arto Anttila, editors. 1995. *Constraint Grammar: A Language-Independent System for Parsing Running Text*. Number 4 in Natural Language Processing. Mouton de Gruyter, Berlin and New York. ISBN 3-11-014179-5.

Kristiina Muhonen and Tanja Purtonen. 2011. Creating a dependency syntactic treebank: Towards intuitive language modeling. In Kim Gerdes, Eva Hajičová, and Leo Wanner, editors, *Proceedings of the International Conference on Dependency Linguistics*, pages 155–164, Barcelona. ISBN 978 84 615 1834 0.

Tommi Pirinen. 2011. Open Source Finnish Morphology (OMorFi). Department of Modern Languages, University of Helsinki. `http://www.ling.helsinki.fi/kieliteknologia/tutkimus/omor/`.

Atro Voutilainen, Krister Lindén, and Tanja Purtonen. 2011. Designing a dependency representation and grammar definition corpus for Finnish. In *Proceedings of III Congreso Internacional de Lingüística de Corpus (CILC 2011)*.