Improving Automatic Identification of Coordinated Ellipsis in Dutch

Martijn W. Hennink
s1158171
27 november 2006

Martijn walks and Martijn whistles.

Jennifer Spenader (Kunstmatige Intelligentie)
Petra Hendriks (Nederlands/KI)
Gosse Bouma (Informatiekunde)

Kunstmatige Intelligentie
Rijksuniversiteit Groningen

ABSTRACT
Ellipsis, the non-expression of sentence elements whose meaning can be retrieved by the hearer, is a common phenomenon in both spoken and written language. This research focuses on three types of ellipsis, namely conjunction reduction, gapping, and right node raising (examples a, b, and c below).

a) Jan koopt appels en (Jan) verkoopt peren.
Jan buys apples and (Jan) sells pears.
b Jan koopt appels en Piet (koopt) peren.
Jan buys apples and Piet (buys) pears.
c Jan koopt (appels) en Piet verkoopt appels.
Jan buys (apples) and Piet sells apples.

Frequency data on ellipsis in Dutch was gathered from a 86,347-word selection of the spoken CGN corpus and a 192,219-word selection of the written Clef corpus, both automatically parsed by the Alpino parser. Initially, 250 conjoined sentences were manually analysed for each corpus. This provided initial frequency data and helped in developing search patterns. Automatic searching was successful for conjunction reduction (b), but right node raising (a) and gapping (c) were parsed incorrectly by Alpino, making the search difficult. An alternative solution involving searching for intransitive parses of typically transitive verbs was used to expand the search for right node raising, the most difficult of the three. The obtained data suggests that the frequency of the different types of ellipsis in Dutch is similar to that in English (Meyer, 2002).

955

CONTENTS

PREFACE

I have always had a great interest in both the quirky tricks language pulls on us and those we pull on language, and long before the time came to pick a topic for my graduation research I had spoken to Petra about my intentions to graduate on research of some interesting linguistic phenomenon. When I finally came to her to ask if she would like to oversee my final project she already had a very interesting topic waiting for me. I am speaking of ellipsis, and it did indeed look very interesting. After some initial brainstorming on the basis of a book by Meyer I gladly accepted the topic and started work on my research thesis. One of the aims was to gather frequency data on ellipsis in Dutch with the help of manual research and automatic research with Alpino.

Alpino, incidentally, was developed at the Rijksuniversiteit Groningen, where I study. Advice and resources were right around the corner, which meant that hopefully my findings could be put to good use too. It may not seem intuitive to use an automatic parser, knowing it is impossible to parse natural languages correctly as long as ellipsis is not fully understood, but the beauty of using Alpino was that I could spot parse flaws and search for ellipsis at the same time. The automatic search proved to be arduous, but in the end the results of this study and the discussion of literature on ellipsis provided results that nicely synched with those of Meyer's research.

I would like to thank Jennifer Spenader, Petra Hendriks, and Gosse Bouma for their support and help throughout this project. I know working with me must have been very demanding at times, but I hope you are pleased with the end result. I know I am. Special thanks go out to Gertjan van Noord as well, for providing the required corpora and use of the Alpino parser. Finally, I can't forget to thank Corien, my girlfriend, and my parents, they never stopped believing in me. It's a blessing to have someone to go to when things are not going according to plan.

ABBREVIATIONS

Ø	Empty (elided) category.
<u>underlined</u>	Antecedent. What the elided category refers to.
( )	The category may be optionally elided.
*	The sentence is ungrammatical.
?	The sentence is of questionable grammaticality.
[ ]	Coordination
SOV	Indicates word order, in this case Subject, Object, Verb. Other word orders include SVO and OVS.
NP	Noun phrase.
PP	Prepositional phrase.
VP	Verb phrase.
CR	Conjunction Reduction, a type of coordinated ellipsis
RNR	Right Node Raising, a type of coordinated ellipsis

1. INTRODUCTION

Most people haven't heard of ellipsis, yet everybody uses it. Whenever you have a conversation, there is a big chance ellipsis is involved. If you pay close attention to what you are saying, you'll probably notice that you're not verbalizing each and every word, yet your listener doesn't seem to have any problem understanding you. In speech, as well as in writing, people tend to omit words all the time. This is what ellipsis is all about.

In his book on corpus linguistics, Meyer (2002) discusses elliptical coordination in some detail. According to Meyer, elliptical coordination is when an element is left out of a sentence without affecting the meaning of the sentence. Consider the following example. 1a shows an uncondensed sentence, whereas 1b shows the same sentence after ellipsis takes place. The empty element "Ø" indicates the site at which a word has been elided. The empty element correlates with the underlined word, which I will refer to as the "antecedent".

Example 1.
a. Maria eats blueberry pie and Maria drinks orange juice.
b. <u>Maria</u> eats blueberry pie and Ø drinks orange juice.

Most people would automatically omit the second instance of *Maria* in sentence 1a, as the sentence feels both artificial and redundant if *Maria* is left intact. Removing the word as in sentence 1b doesn't have an effect on the meaning of the sentence (though see chapter 2-1, What Is Ellipsis?) and is indeed a perfect example of syntactic ellipsis. It also showcases one of the main reasons to use ellipsis, namely speaker's economy of effort[1]. This principle states that the speaker will use only as much effort as is needed to convey the intended message. (More reasons follow in chapter 2-2, Why Use Ellipsis?)

Ellipsis encompasses a wide variety of linguistic phenomena, all sharing a very basic principle, namely a missing element that can be retrieved by the hearer or reader. The variety is so wide in fact, that I can't possibly cover all of it within the scope of my research. At the end of this paper I will discuss some other types of ellipsis, but to keep things manageable I will focus on only three types of ellipsis, namely right node raising, conjunction reduction, and gapping. These are essentially the same three types used in Meyer's research, and they occur in Dutch as well, thus making a comparison possible. Below is an example sentence for each of these types.

Example 2.
a. Right node raising: Mary loves Ø and Joe hates <u>turkey</u>.
b. Conjunction reduction: <u>Mary</u> loves turkey and Ø hates chicken.
c. Gapping: Mary <u>loves</u> turkey and Joe Ø chicken.

Meyer's study of ellipsis in English (Meyer, 1995) produced some very interesting results. He used a 96,000-word corpus, based on sections of the Brown Corpus and the American component of the International Corpus of English (ICE). The corpus contained both different types of writing and speech. Instead of the usual ellipsis categories, he adopts Sanders' linear system (Sanders, 1977), which identifies ellipsis based on the site where it takes place. Since Meyer's research focuses only on English however, we can substitute these terms with the regular ones[2]. Conjunction reduction was by far the most prevalent form, accounting for 86

[1] Speaker's economy of effort is based on Zipf's principle of least effort (Zipf, 1949) and his Force of Unification.
[2] In English, C-ellipsis corresponds to right node raising, D-ellipsis to conjunction reduction, and E-ellipsis to gapping. Note that Sander's linear system deviates from the view that the site of these three types of ellipsis shift in languages with a surface structure word order that differs from English (see chapter 2-3 and further).

percent of the elliptical coordinations identified in the corpus. Right node raising and gapping occurred only in 2 percent and 5.5 percent respectively. The remaining 6.5 percent consists of constructions with more than one type of ellipsis.

Also interesting are the differences in frequency of ellipsis between writing and speech. In speech only 40 percent of the cases where elliptical coordination was possible actually used the reduced form. In writing, however, ellipsis occurred 73 percent of the time when it was possible. Meyer provides two possible explanations for this difference, natural flow of speech and semantically less dense discourse on the listener's part. The natural flow argument seems to be based on the idea that a speaker starts with forming the uncondensed sentence in his head. Adding in ellipsis might be elegant, but it is an extra processing step on the speaker's part, one which is often skipped in a conversation. In other words, because forming ellipsis requires more effort on the speaker's part, often he will simply utter the uncondensed sentence. Notice that this theory assumes the opposite of the rule of Speaker's Economy of Effort mentioned earlier, which is grounded in the belief that adding ellipsis decreases effort. The high percentage of fully pronounced sentences means the natural flow of speech theory can't be simply discarded though.

The argument of semantically less dense discourse is based on the idea that a speaker wants to get his message across. Ellipsis requires more processing by the hearer and is thus semantically more dense. By providing a sentence without ellipsis he ensures that the listener has an easier time interpreting the meaning of the sentence, which in turn speeds up conversation.

It is not clear, however, that Meyer's arguments are sufficient to explain the lower frequency of ellipsis in speech. There are also factors that seem to increase the viability of ellipsis, which, I would expect, could counter the effect of the natural flow of speech and the semantically less dense discourse arguments. For one, speech provides much more contextual information than writing. Speakers gesticulate, from pointing out the object of the conversation to stressing parts of speech. Speakers might share a background, where common information is silently understood, all adding to the redundancy of speech and making conversation more eligible to ellipsis. My study does not focus on these questions, but it would make for interesting future research (see also chapter 6.4, Future Research).

1-1. Frequency Analysis

Ellipsis forms a crucial part of natural language, but up until now hardly anyone has tried to tackle ellipsis in terms of frequency in Dutch. Hoeksema (to appear) did gather examples of gapping in Dutch, but since these weren't found by a systematic search in real texts but simply noted when found, it is impossible to calculate frequency data from it. There has also been a good deal of theoretical research on the semantic level, and frequency analysis could confirm these theories. Thus, corpus analysis can provide a solid basis for further research. Also, with the ultimate intention of a 'perfect parser' in mind, it is important to note that ellipsis varies greatly between languages, both in usage and types of ellipsis. Each language has its own characteristics and therefore needs to be individually subjected to corpus analysis.

The goal of this study, therefore, is to provide frequency data on ellipsis in Dutch using Dutch corpora, and to compare the results with Meyer's observations for English if possible. Because it is not feasible to gather all this information by hand, I use the Alpino parser (Bouma et al., 2001) to find examples of the aforementioned three types of syntactic ellipsis, as it's one of the best parsers of natural language available right now for Dutch.

1-2. Overview

Chapter two outlines the theoretical framework of my research. In it I explain what ellipsis is, and discuss theories about why we use it. In this chapter I also discuss right node raising, conjunction reduction, and gapping in greater detail. I focus on finding good definitions for each of the three types, that will allow them to be reliably identified for Dutch. Getting the definitions right is important for the construction of search patterns and for determining whether ellipsis is handled correctly by the automatic parser.

Chapter three elaborates on the two corpora, on Alpino and on the search methods I used in my exploration of ellipsis. I focus on the dependency structures Alpino assigns to its parses, since that is the feature allowing me to use XML-queries to locate certain characteristics of syntactic ellipsis.

Chapter four explains the steps I took while trying to find ellipsis with Alpino. I give examples of problematic situations and determine what kind of search patterns would be possible for each of the three ellipsis types, based on the way the corpora are parsed.

Chapter five goes over the results of the automatic search for ellipsis, which then will be discussed in chapter six. I will present some possible solutions to enhance future parsing of syntactic ellipsis, as well as discuss other types of ellipsis that must be covered in future work in chapter six as well. Finally, this paper ends with a conclusion to highlight the major findings of my research.

2. THEORETICAL FRAMEWORK

2-1. What Is Ellipsis?

Ellipsis is a very common phenomenon in English, as well as in Dutch[3]. There are a lot of linguistic phenomena labelled under ellipsis though, and the domain needs to be delimited to be able to say anything relevant on the topic. So, the first question I will try to answer is what exactly *is* ellipsis? Meyer (2002) defines ellipsis as a coordination in which some element is left out of a sentence without affecting the meaning of the sentence. Ellipsis however does not necessarily happen in coordinated sentences alone and, as we will see later on, it *can* change the meaning of an expression. Hendriks and Spenader (2005) define ellipsis as the non-expression of sentence elements whose meaning can be retrieved by the hearer. Their definition rightly takes into account that deletion can also take place in non-coordinated sentences. Also, it avoids stating that the meaning of an elided sentence doesn't change, so I will adopt this definition.

According to our definition ellipsis is retrievable from context. Though retrieving it does not necessarily take place on a syntactic level, hopefully a parser will be able to find a syntactic clue to restore and correctly parse a sentence with ellipsis. Interestingly, an ongoing debate in theoretical linguistic work revolves around what the role of syntax in the representation of ellipsis is (Kennedy, 2001). One approach believes that elided material has syntactic structure at some level of representation, with the grammar containing a means of blocking the pronunciation of the elided material in the surface form. The other approach rejects this view and recovery of meanings from context is enough to resolve ellipsis (i.e. syntax is not necessary for this). Though the problem hasn't been solved, the whole idea of using an automatic parser to find ellipsis in a corpus implies a belief that, at some level, ellipsis is represented in, or at least signalled by, the syntax.

An example of a sentence where context possibly enables recognition of ellipsis is given below. I will follow the structure of example 3 in the rest of this paper. That is, the source sentence is followed by a gloss, which is in turn followed by the translation into English. If the example is taken from an external source I will cite the source between brackets, right after the example number.

Example 3.
Ik <u>studeer</u> kunstmatige intelligentie en Tim Ø economie.
I <u>study</u> artificial intelligence and Tim Ø economy.
I study artificial intelligence and Tim economics.

Here the verb *study* is elided in the second clause under semantic identity with the same verb in the first clause. That is, the empty element correlates with the antecedent *study*. This form of ellipsis is called *gapping*, a fairly common type in both English and Dutch. Notice that in this particular example, ellipsis is only possible because of the coordination present in the sentence. This observation holds true with gapping in general, and applies to conjunction reduction and right node raising as well (Hudson, 1976; Van Oirsouw, 1984). Since I will refer to these forms of ellipsis frequently in the rest of this paper I will present a few Dutch examples (with translations) for the reader's reference, with <u>the antecedent</u> of the elided part underlined.

[3] Based on my manual study, approximately 20% of all coordinated sentences in written Dutch features ellipsis or could feature ellipsis but was fully pronounced (50 and 3 out of 250). For spoken Dutch this figure also nears 20% (29 and 14 out of 250).

Example 4.

a Tim koopt Ø en Ben eet <u>een appel</u>. *[Right-node Raising (RNR)]*
Tim buys Ø and Ben eats <u>an apple</u>.

b. <u>Tim</u> koopt een appel en Ø eet een banaan. *[Conjunction Reduction (CR)]*
<u>Tim</u> buys an apple and Ø eats a banana.

c. <u>Tim</u> loopt en Ø luistert naar muziek. *[Conjunction Reduction (CR)]*
<u>Tim</u> walks and Ø listens to music.

d. Tim <u>koopt</u> een appel en Ben Ø een banaan. *[Gapping]*
Tim <u>buys</u> an apple and Ben Ø a banana.

Note that conjunction reduction doesn't require the clause-final position of the preceding conjunct and the clause-medial position of the following conjunct to be filled, as is the case in the second example above. Quite commonly, two types of ellipsis can occur in the same sentence. Conjunction reduction can occur together with either right node raising or gapping, as in the following sentences.

Example 5.

CR + RNR: <u>Tim</u>$_1$ koopt Ø$_2$ en Ø$_1$ eet <u>een appel</u>$_2$.
<u>Tim</u>$_1$ buys Ø$_2$ and Ø$_1$ eats <u>an apple</u>$_2$.

CR + Gapping: <u>Tim</u>$_1$ <u>koopt</u>$_2$ een appel en Ø$_1$ Ø$_2$ een banaan.
<u>Tim</u>$_1$ <u>buys</u>$_2$ an apple and Ø$_1$ Ø$_2$ a banana.

Normally it is easier to interpret these cases of multiple ellipsis as though they were normal coordinations of two VPs (e.g. "koopt en eet") or NPs (e.g. "een appel en een banaan"). This avoids the use of ellipsis, which is, in terms of rules, more complex than simple coordination in these cases[4]. Also, the line between a combination of two types of coordinated ellipsis and a simple list of NPs is really thin and hard to draw. For example, consider the following sentence from the Clef-corpus. Sentence 6a is taken directly from the corpus, and 6b attempts to reconstruct the sentence as if 6a features ellipsis. The reconstructed parts are between brackets, indicating that they should be optionally elidable.

Example 6.

a. Naast Dennis Bergkamp en Wim Jonk heeft Inter nog drie andere buitenlanders.
Besides Dennis Bergkamp and Wim Jonk has Inter also three other foreigners.
Besides Dennis Bergkamp and Wim Jonk, Inter also has three other foreigners.

b. Naast Dennis Berkamp (heeft Inter nog drie andere buitenlanders) en (naast) Wim Jonk heeft Inter nog drie andere buitenlanders.
Besides Dennis Bergkamp (has Inter also three other foreigners) and (besides) Wim Jonk has Inter also three other foreigners.
Besides Dennis Bergkamp, Inter also has three other foreigners and besides Wim Jonk, Inter also has three other foreigners.

As you can see, it is possible to reconstruct an elided part by treating 6a as a combination of CR and gapping, but it makes for a very messy read and it is not quite clear anymore what the meaning of the sentence is. One could argue that one of the three other foreigners in Bergkamp's case is Jonk and vice versa. There are also coordinations where it is flat-out impossible to reconstruct an elided part without altering the sentence. In the following example the word *elk* (*each*) prohibits a grammatical reconstruction, as indicated by the *.

[4] Occam's razor comes to mind. When two rules are equally adept at explaining a phenomenon, choose the least complex one. Obviously Occam's razor doesn't always apply, but in this case I think it does.

Example 7.

a. De Kenyaan Simon Chemwoyio en de Ethiopiër Fita Bayesa werden elk 12.000 gulden rijker.
The Kenyan Simon Chemwoyio and the Ethiopian Fita Bayesa became each 12,000 gulden richer.
The Kenyan Simon Chemwoyio and the Ethiopian Fita Bayesa each earned 12,000 gulden.

b. De Kenyaan Simon Chemwoyio *(werd 12.000 gulden rijker) en de Ethiopiër Fita Bayesa werden elk 12.000 gulden rijker.
*The Kenyan Simon Chemwoyio *(became 12,000 gulden richer) and the Ethiopian Fita Bayesa became each 12,000 gulden richer.*
*The Kenyan Simon Chemwoyio earned 12,000 gulden and the Ethiopian Fita Bayesa each earned 12,000 gulden

Because of the complications that arise when trying to treat certain coordinates as a case of ellipsis and when trying to determine whether or not it is actually a combination of two types of ellipsis, I will not address these cases in my research.

2-2. Why Use Ellipsis?

Now we know what Ellipsis is, but why would one want to use it? I already mentioned that ellipsis can be used to remove repetition of certain elements in a sentence. There is little advantage to including superfluous information, so it's easier on the speaker's part to just leave it out. This is called the principle of Speaker's Economy, widely acknowledged to be one of the driving principles behind many linguistic phenomena, including ellipsis. Hendriks and Spenader (2005) discuss a number of additional purposes of ellipsis, which shed further light on my question.

According to Hendriks & Spenader, ellipsis can remove readings, and thereby clarify the meaning of a sentence (they took the *b*-sentence in the following example from Partee & Rooth (1983)).

Example 8. (Hendriks & Spenader, 2005)

a. A fish walked and a fish talked.
b. A fish walked and Ø talked.

The elided sentence has a different meaning from the complete one. Sentence *a* has two different readings, one of which assumes one fish walked and talked, the other assuming the walking and talking is performed by two different fish. By eliding the subject in the following conjunct, the second reading has been rendered impossible, obviously narrowing down the ambiguity in the sentence.

Their next paragraph covers conveying non-expressible aspects of meaning. They argue that an elided sentence element need not necessarily be expressible. Consider their first example.

Example 9. (Carlson, 1977)

a. Wolves get bigger Ø as you go north from here.
b. *Wolves get bigger than ??? as you go north from here.

Clearly, the intended reading is not that a particular wolf would get bigger as it migrates northwards, but rather that separate populations of wolves are bigger in size, the further north they live. Try to fill in the question marks in sentence 9b, and you'll find that it's impossible to express this in a way that feels right. You would have to use a comparative relation

between the same referent (as in 9c below), and that's exactly what, according to Hendriks and Spenader, is restricted.

c. *Wolves get bigger than wolves as you go north from here.

Establishing discourse coherence also plays an important role in the why of ellipsis. By eliding certain elements, the speaker can add to the flow of text.

Example 10. (Hendriks & Spenader, 2005)

a. John walked. John talked.
b. John walked. He talked.
c. John walked and Ø talked.

Here we can see that using ellipsis not only enhances the flow of a conversation, but also eliminates ambiguity. Whereas there could be two different Johns in sentence 10a, and even, with some gesticulation, two different subjects in 10b, no such interpretation is possible in 10c. The missing subject must be the same as the subject of "walked".

The last use of ellipsis Hendriks & Spenader cover is that of establishing a positive relationship with the reader. Often people insert blanks into their speech or writing for the hearer or reader to fill in. Sometimes this is done as a means to show camaraderie, at other times simply to "short talk". A few good examples are given in the article, but it's quite easy to think up some of yourself, since it is indeed a common use of ellipsis.

Example 11. (a: Hendriks & Spenader, 2005)

a. If your husband routinely comes home late with lipstick on his collar... (then he must be having an affair)
b. (Are) you from here?
c. (Have you) seen anything of interest?
d. This is your last chance, next time... (I won't be as forgiving)

Even though there are many reasons to use ellipsis, some types of ellipsis are used more often than others. This seems to indicate that there are certain constraints at work, constraints that don't affect all types in equal measure. Sanders (1977, in Eckman) argues that the difference in prevalence can be explained by two processes, the *suspense effect* and the *serial position effect*.

The suspense effect predicts that ellipsis will be relatively undesirable if the site of ellipsis precedes the antecedent of ellipsis, since the suspense created by the anticipation of the elided item places a processing burden on the hearer or reader (Meyer, 2002). Thus in example 2a the hearer must wait until the very end of the sentence to find out what Mary loves. This rule predicts right node raising is undesirable because there the elided category precedes its antecedent.

The serial position effect is based on research demonstrating that when given memory tests, subjects will remember items placed in certain positions in a series better than other positions. Therefore, the closer an antecedent is to the start (or ending) of a sentence, the easier it is to remember it when the reader gets to the site of ellipsis. This rule favours conjunction reduction, as in example 2b, as its antecedent heads off the whole sentence. It is easy to see that this effect would once again predict that right node raising is the most undesirable form of the three types I am studying.

With these two restrictions in mind, one would expect conjunction reduction to be the most prevalent because it doesn't violate the suspense effect and adheres optimally to the serial position effect. Right node raising should be the least prevalent, because it both violates the suspense effect and does badly on the serial position effect, since the antecedent is in the middle of the sentence. These expectations are nicely reflected in Meyer's frequency data, thus the existence of these two processes seem to be likely. The question is if the same effects will play out in Dutch. In the next few sections I will look at each of the three types of ellipsis I studied and look at the theory behind each one.

2-3. Conjunction Reduction (CR)

Conjunction is the ellipsis of a subject in a coordinated sentence, as in example 4b and 4c, the first of which I repeat below.

Example 12.
Tim koopt een appel en Ø eet een banaan.
Tim buys an apple and Ø eats a banana.
Tim buys an apple and eats a banana.

Of the three ellipsis types I am looking for, this one is the easiest. First, the antecedent precedes the site of ellipsis, and second, conjunction reduction leaves behind chunks that can easily be coordinated. Example 12 shows both of these observations. In 12, the second instance of the subject *Tim* is elided, and the ellipsis of the subject leaves behind a verb and its object on each side of the coordinator *and*. Besides being the easiest type of the three, conjunction reduction also is the most common. Meyer (1995) found that the type of ellipsis in English coordinations was conjunction reduction 79% percent, and partial conjunction reduction (in which case the subject is replaced by a pronomen) 7% of the time. I fully expect Dutch to behave in the same way.

> Conjunction Reduction deletes the subject in a coordinate separate from that containing its antecedent. The conjuncts must exhibit a parallel syntactic structure and conjunction reduction can only apply forward, that is, the antecedent must precede the site of ellipsis.

Note that the definition above doesn't restrict the elided subject to be in the first position of the following conjunct. There's a good reason for this, as we'll see in the next section.

2-4. SGF-Coordination (SGF)

SGF-coordination, also sometimes referred to as subject gapping, stands for Subject Gap in Fronted/Finite clause coordination (Höhle, 1983 and Hendriks, 2004). It's a kind of ellipsis that occurs quite often in Dutch and in German, but only seldom in English. Below is an example for both Dutch and English, since SGF happens under different conditions in both languages.

Example 13a. (Clef, adapted)
Tegen Napoli draaide ik weg van mijn opponent en nam Ø de bal aan.
Against Napoli turned I away from my opponent and took Ø the ball on.
Against Napoli I turned away from my opponent and collected the ball.

Example 13b. (Harbusch & Kempen, 2006)
Why did you leave but didn't Ø warn me?
Why did you leave but didn't warn me?.

As far as I know, example 13b is the only way to obtain SGF in English. In Dutch, PP's and can be fronted, forcing the VP to move to second position. As long as parallelism is retained, subject ellipsis under syntactic identity is still possible, resulting in sentences like the one in example 13a. The question that has to be answered is whether or not SGF should be counted as a case of conjunction reduction. If I do I will need two separate search patterns, since SGF-coordination differs syntactically and semantically from traditional CR. If I don't, I leave an important case of coordination ellipsis (I found it to be just as common as gapping) out of the picture.

To answer this dilemma I asked myself if CR and SGF were really that different. Both are restricted to coordinations and elide identical subjects in the following conjunct. What if there underlying structure is the same? It is possible that the rule for PP fronting applies after the rule for conjunction reduction. This would mean that both (presumably – see 2-6 Gapping) have an underlying SOV word order at the time of ellipsis and the word order is determined after that. To my knowledge it is not possible in Dutch to check the order in which these two rules apply, but since this scenario is possible, I will treat SGF as a special case of CR, and as such will have to fabricate a unique search pattern for it.

2-5. *And Then* Coordination

The definition on CR also requires that the conjuncts must exhibit a parallel structure. This requirement is of vital importance to all three types of coordinative ellipsis, and this section presents an example that makes this absolutely clear.

There exists a difference between Dutch and English coordination regarding the *and then* coordination. This coordinator can be translated into Dutch as either *en dan*, used in present and future tense, or *en toen*, used in past tense. A quick example should make things clear[5].

Example 14a. (Dutch)
Marie liep naar de winkel en toen reed ze naar huis.
Marie walked to the store and then drove she to home.
Marie walked to the store and then she drove home.

Example 14b. (English)
<u>Marie</u> walked to the store and then Ø drove home.
Marie walked to the store and then drove home.

Immediately we see a difference. In sentence 14a ellipsis is prohibited, whereas in example 14b conjunction reduction is allowed. This difference seems to stem from the fact that the *and then* coordination apparently invokes the V2 rule (verb second) in Dutch. This rule forces the verb to be in second place, which changes the relative word order of subject, verb, and object in, for example, sentences with a fronted PP. Because the *and then* coordinator forces verb movement only in the following conjunct, but at the same time does not elicit verb movement in the preceding conjunct, the parallelism disappears and ellipsis (be it CR, RNR, or gapping) is no longer possible. Thus, this case proves that the V2 rule causes less parallelism in Dutch, which leads to restriction on ellipsis. More importantly however, this observation alone makes clear that parallelism is a necessary prerequisite for coordinative ellipsis.

[5] Gosse Bouma rightly noted that, by making "toen" an adverb, ellipsis is possible in Dutch, as in "Marie liep naar de winkel en (ze) reed toen naar huis." This doesn't make the discussion irrelevant though. Also, Dutch subordinating conjunctions, like "omdat" ("because") and "tenzij" ("unless"), force an SOV word order in the following conjunct while retaining the SVO order in the preceding conjunct, which would be an analogue case that does always render (coordinated) ellipsis impossible.

2-6. Gapping

Gapping refers to the ellipsis of a verb in a coordinated sentence, as per example 4d, which is repeated and expanded below.

Example 15.
Tim koopt een appel en Ben Ø een banaan.
Tim buys an apple and Ben Ø a banana.
Tim buys an apple and Ben a banana.

The literature defines gapping in a number of different ways, a few of which I will discuss here. Below I list a number of statements, definitions to a certain extent, in chronological order. Commentary inside the quotes appears between square brackets.

a. "Note that Gapping operates only forward in English – that is, in *n* conjoined sentences, it is the leftmost occurrence of the identical main verb that causes the *n-1* following occurrences to be deleted. In Japanese, an SOV language, exactly the posite opis (sic.) [opposite is] the case – it is the rightmost verb among *n* identical verbs that is retained." *(Ross. 1970)*

b. "The simplest cases of Gapping delete the verb of one or more clauses conjoined to the right of a clause containing the same verb..." *(Jackendoff, 1971)*

c. "Gapping is an ellipsis rule that applies in coordinate structures to delete all but two major constituents from the right conjunct under identity with corresponding parts of the left conjunct..." *(Hankamer & Sag, 1976)*

d. "Consider the rules of Gapping [...] in Dutch. Gapping deletes verbs [...] under identity in coordinate structures." *(Van Oirsouw, 1984)*

e. "In clausal coordination, it seems that we most often find analipsis [*i.e.* forward gapping] of a constituent in the second coordinand. If the constituent is in a clause-medial position (thus leaving a gap), this type of analipsis is called gapping..." *(Haspelmath, 2004)*

f. "Jackendoff (1971 as cited in Lobeck, 1995) outlined 4 differences between gap and ellipsis[6]:

1. A gap must be flanked by lexical material. An ellipsis can be phrase-final.

2. A gap must occur in a coordinate, but not subordinate (adjunct or complement) clause separate from that containing its antecedent. An ellipsis can occur in a coordinate or subordinate clause separate from that containing its antecedent.

3. A gap cannot precede its antecedent. An ellipsis can precede its antecedent under certain conditions.

4. A gap need not be a phrase. An ellipsis must be a phrase.

[6] Note that Jackendoff means VP-ellipsis here, not ellipsis in general. VP-ellipsis is the elision of a verb phrase. E.g. "John ate lunch and we did (ate lunch) too."

The above examples suggest that gapping can operate on a phrasal constituent, but is not required to. Rather, the fundamental element for a well-formed gap is the presence of flanking material, which appears to play no crucial role in the process of forming a verb phrase (VP) ellipsis..." *(Hansen, 2005)*

These statements contain a lot of contradictory information, most notably on whether the elided object must be (or contain) a verb and on the position in which gapping takes place. They all agree that gapping takes place only in coordinated sentences, confirming the claims of Hudson (1976) and Van Oirsouw (1984). Ross also claims that English, as a SVO language, gaps forward, and Japanese, as a SOV language, gaps backward. Once again, the material between brackets is optionally subject to ellipsis.

Example 16. (adapted from Ross, 1970)
I ate fish, Bill (ate) rice, and Harry (ate) roast beef.
Tom has a pistol, and Dick (has) a sword.
I want to try to begin to write a novel, and Mary (wants (to try (to begin (to write)))) a play.

Example 17. (adapted from Ross, 1970)
Watakusi was akana o (tabat), Biru wa gohan o tabeta. (sic.)
Watakusi was sakana o (tabate), Biru wa gohan o tabeta.
I {prt} fish {prt} (ate), Bill {prt} rice {prt} ate.
I ate fish, and Bill (ate) rice.

This definition overlaps with the definition of right node raising in that it elides a verb in the right-most position of the preceding conjunct. It seems however that Ross is onto something. A changed word order would predict a change in the site of gapping, right node raising and conjunction reduction, since those types of ellipsis are site-bound, and thus depend on the word order. I gather that the basic premises of the above statements, like "gapping can only apply in a coordinated sentence", still apply, but that the English-specific rules (or more generally, those specific to SVO-class languages) need to be dropped. This view renders the definitions in *c* and *e*, and rule 1 of statement *f* plus the conclusion of statement *f* moot, and counters the claims that ellipsis can only apply forward (or in the right-node of a conjunction) of statement *b* and rule 3 of statement *f*, because each of these statements depends on a rigid word order. That is, the flanking material discussed in the conclusion of statement *f* only flanks in an SVO language like English. In Japanese, for example, the SOV word order obviously prevents flanking of the V-component. Likewise, all rules concerned with a fixed site of gapping fail to accommodate for a shifted word order that elicits gapping at another site. Therefore, in defining gapping I will use the definitions (Ross (1970), Van Oirsouw (1984) and partly Jackendoff (1971 as cited in Lobeck, 1995) that can be extended to account for gapping in non-SVO languages.

Gapping deletes verbs in a coordinate, but not subordinate (adjunct or complement) clause separate from the clause containing its antecedent. The conjuncts must exhibit a parallel syntactic structure. In sentences with a verb-final surface word order gapping can operate either backwards or forwards, otherwise it must operate forwards.

There are several languages that have an underlying SOV word order but can produce sentences with another surface structure. One of those languages, as Koster (1975) argues convincingly, is Dutch. Ross (1970) however claims that Dutch has an underlying SVO word

order, just like English. Either way, one needs to explain how Dutch forms sentences with a surface word order different from its underlying word order. Observe:

Example 18a. (main clause, SVO)
Marie koopt een boek.
Mary buys a book.
Mary buys a book.

Example 18b. (subordinate clause, SOV)
...dat Marie een boek koopt.
...that Mary a book buys.
...that Mary buys a book.

So, Koster will have to explain how the main clause in 18a is formed from an underlying SOV word order, and Ross has to explain how an underlying SVO word order leads to the subordinate clause in 18b. They both use a rule of verb movement, however, Koster's rule is much more elegant and simple at shifting an underlying SOV form to an SVO surface structure than Ross's is at doing the reverse. In addition, Koster gives several examples of Dutch phenomena that can be easily explained through the single rule of verb movement, where Ross needs a rule of particle movement (a difficult one in Dutch) as well.

In light of Koster's arguments, I will assume Dutch has an underlying SOV form, though it's only in subordinate clauses that Dutch produces SOV surface structures. The rule of verb movement forces an SVO surface structure in main clauses, and enables a VSO surface word order in questions (as in English questions). Examples 19 through 21 display each of these surface structures with an attempt to gap both forwards and backwards on each word order.

Example 19a. (SVO, forward gapping)
Marie <u>koopt</u> een boek en Jan Ø een strip.
Marie <u>buys</u> a book and Jan Ø a comic.
Marie buys a book and Jan a comic.

Example 19b. (SVO, backward gapping)
*Marie Ø een boek en Jan <u>koopt</u> een strip.
Marie Ø a book and Jan <u>buys</u> a comic.
Marie buys a book and Jan a comic.

Example 20a. (SOV, forward gapping)
Ik weet dat Marie een boek <u>koopt</u> en Jan een strip Ø.
I know that Marie a book <u>buys</u> and Jan a comic Ø.
I know Marie buys a book and Jan a comic.

Example 20b. (SOV, backward gapping)
Ik weet dat Marie een boek Ø en Jan een strip <u>koopt</u>.
I know that Marie a book Ø and Jan a comic <u>buys</u>.
I know Marie buys a book and Jan a comic.

Note that though examples 20a and 20b features no flanking material, they still adhere to the same rules (aside from the site of ellipsis) as the gapping example in 19a. English doesn't have SOV sentences though, so it seems that early research simply drew the false conclusion because it focused too much on English. Likewise, question sentences like in example 21 below didn't get the attention of research on gapping either and missed out on the gapping label as well, leaving the prerequisite of flanking material intact.

Example 21a. (VSO, forward gapping)
Koopt Marie een boek en Ø Jan een strip?
Buys Marie a book and Ø Jan a comic?
Is Marie buying a book and Jan a comic?

Example 21b. (VSO, backward gapping)
*Ø Marie een boek en koopt Jan een strip?
Ø Marie a book and buys Jan a comic?
Is Marie buying a book and Jan a comic?

As we can see, our definition holds true. Only the sentence with a verb-final surface structure allows for backward gapping (20b), whereas all of the examples with forward gapping are permitted. There are however two other kinds of ellipsis that might falsely fall under the current definition, VP-ellipsis and pseudogapping. Hoeksema (to appear) gives the following sentences for comparison. For clarity I marked the antecedent and the site of ellipsis.

Example 22. (adapted from Hoeksema, to appear)
a. *Pseudogapping:* That may not bother you, but it does Ø me.
b. *Gapping:* Smoke bothers Fred, and loud music Ø Fred's parents.
c. *VP-ellipsis:* Smoke might have bothered Fred, but it didn't Ø.

As Hoeksema notes, pseudogapping resembles gapping in that it elides a verb (plus additional elements), while nonverbal elements like direct objects may be left behind as remnants. Like VP-ellipsis, pseudogapping leaves behind an auxiliary verb. There are however a few major differences that set pseudogapping and VP-ellipsis apart from gapping. First, example 22b features parallelism. Indeed, Féry and Hartmann (2005) state, in regards to right node raising and gapping, that "the conjuncts must exhibit a parallel syntactic [...] structure". VP-ellipsis and pseudogapping seem to need contrast between the two conjuncts though.

The second important difference can be seen when we try to apply these types of verb-ellipsis to comparative clauses.

Example 23.
a. We like cats more than they do Ø dogs. *[pseudogapping]*
b. *We like cats more than they Ø dogs. *[gapping]*
c. We like cats more than they do Ø. *[VP-ellipsis]*

Notably, gapping is the only type not allowed here, which was to be expected if we remember Hudson's (1976) and Van Oirsouw's (1984) claim that gapping can only take place in a coordinate clause. Last but not least, Levin (1985) notes that VP-ellipsis can operate backwards in English, which is a direct violation of my gapping rule, indicating that VP-ellipsis and gapping are indeed not the same.

Example 24. (Levin, 1985)
Although it doesn't always Ø, it sometimes takes a long time to clean the hamster's cage.

Still, the above evidence seems to imply that at least some instances of pseudogapping and VP-ellipsis fall under the above definition of gapping, which leads me to a slightly expanded new definition which, most notably, includes Féry's and Hartmann's notion of parallelism.

> Gapping deletes verbs in a coordinate, but *not* subordinate (adjunct or complement) or comparative clause separate from that containing its antecedent. The conjuncts must exhibit a parallel syntactic structure. In sentences with a verb-final surface structure gapping can operate either backward or forward, otherwise it must operate forward.

This definition, finally, seems to reflect gapping quite well, both in English *and* Dutch.

2-7. Right Node Raising (RNR)

Right node raising refers to backward ellipsis at the right-most periphery of the preceding conjunct, as shown in example 4a, which I will repeat (and expand) below for the sake of clarity.

> ***Example 25.***
> Tim koopt Ø en Ben eet <u>een appel</u>.
> *Tim buys Ø and Ben eats <u>an apple</u>.*
> Tim buys and Ben eats an apple.

RNR was originally named so because, in a sentence like the one above, some common element has been raised out of two conjuncts and attached to the right of both of them (Postal, 1974), in the case of example 25 that would be "een appel". Both Postal and Dougherty (1970) agree that RNR is something different from conjunction reduction, but some linguists have argued otherwise. Hudson (1976) settles this dispute by outlining a number of facts that seem to set RNR apart from CR. Apart from the obvious difference in the site of ellipsis, Hudson notices another very important difference, namely that CR is restricted to coordinations, but RNR isn't. Observe (as usual I have underlined the antecedent and marked the site of ellipsis).

> ***Example 26. (a & b from Hudson, 1976 – c from Yatabe, 2001)***
> *a.* I'd have said he was sitting on the edge of Ø rather than in the middle of <u>the puddle</u>.
> *b.* It's interesting to compare the people who like Ø with the people who dislike <u>the power of the big unions</u>.
> *c.* independence of local Ø from central <u>government</u>.

As can be seen in all the above examples, RNR explicitly needs transitive verbs on both sides of the coordinate, whereas CR can also occur in sentences with intransitive verbs. It is also noteworthy that the first example of CR identified as such (Ross, 1967) is in fact a case of RNR!

> ***Example 27. (Ross, 1967)***
> Sally might be Ø, and everyone believes Sheila definitely is, <u>pregnant</u>.

Now that we've established the difference between CR and RNR it is time to find a definition for the latter and to determine the features RNR must abide by. From the rule of gapping presented in the previous paragraph we can infer that verbs cannot be right-node raised, since verbs are elided as a result of gapping. However, consider the following Dutch examples:

Example 28.

a. Tim gaat in Groningen Ø en Ben gaat in Zwolle winkelen.
Tim goes in Groningen Ø and Ben goes in Zwolle to-shop.
Tim is going to shop in Groningen and Ben is going to shop in Zwolle.

b. *Tim gaat Ø in Groningen en Ben gaat winkelen in Zwolle.
Tim goes Ø in Groningen and Ben goes to-shop in Zwolle.
Tim is going to shop in Groningen and Ben is going to shop in Zwolle.

c. *Tim $Ø_1$ in Groningen $Ø_2$ en Ben $gaat_1$ in Zwolle $winkelen_2$.
Tim $Ø_1$ in Groningen $Ø_2$ and Ben $goes_1$ in Zwolle $to\text{-}shop_2$.
Tim is going to shop in Groningen and Ben is going to shop in Zwolle.

d. Tim $gaat_1$ in Groningen $winkelen_2$ en Ben $Ø_1$ in Zwolle $Ø_2$.
Tim $goes_1$ in Groningen $to\text{-}shop_2$ and Ben $Ø_1$ in Zwolle $Ø_2$.
Tim is going to shop in Groningen and Ben is going to shop in Zwolle.

At first sight it seems RNR is meddling with gapping in sentence 28a, but what we see here is a typical Dutch phenomenon. If a "verb-block" is followed by a prepositional phrase[7] (PP), the PP can move into the block to form a verb-final surface structure. Sentences 28b and 28c show that it is impossible to elide the non-final verbs. Even though the surface word order in Dutch main clauses is SVO, the PP movement elicits gapping of those verbs that end up in a clause-final position.

Of course, all the original articles on RNR were written with English in mind. What if, as with gapping, the underlying word order changes? Yatabe (2001) has researched just that in his study on left-node raising (hereafter LNR) in Japanese. LNR in Japanese is almost an analogue mirror function of RNR in English, so much so that it can be classified as such if one were to accept that RNR is not, in fact, limited to ellipsis of the final element of the preceding conjunct. Note how I say element, since it is impossible to be sure this element is indeed an object, unless one were talking about a language with a rigid underlying SVO word order like English. This strengthens my belief that gapping, RNR, and CR are category-bound types of coordinated ellipsis, and not site-bound as Sanders (1977) and hence Meyer (1995) believes. Instead, the site of ellipsis can vary between languages or even clause types, because the site depends solely on surface word order.

> Right node raising deletes objects in a coordinate or comparative clause separate from that containing its antecedent. The conjuncts/comparatives must exhibit a parallel syntactic structure. Right node raising operates backwards, except in object-fronted sentences.

The clean definition of RNR, CR, and gapping as category-bound types of coordinated ellipsis means we can eliminate certain clause constructions that seem to be based on a site-bound idea of ellipsis. The following example is one of those sentences that doesn't contain (pure) RNR according to the definition above.

[7] A prepositional phrase is a sentence element composed of a preposition and usually a complement such as a noun phrase. E.g. (where the PP is italicised) "He was walking *in the woods*", or "He is a student *of physics*."

Example 29 (Clef)[8].
We willen Ø en we moeten marktaandeel winnen.
We want Ø and we must market-share win.
We want to and we need to raise our market-share.

This is a combination of verb-final gapping, which in turn elicits RNR, resulting in a combination of two types of coordinated ellipsis. As said before, the aim of this research is not to find combined ellipsis, but only RNR, CR, and gapping in their pure form, so this type of construction has not been included in the search. It does show the kind of variation possible, and the particular problems a site-based definition might have at identifying certain types of ellipsis in a language with underlying SOV word order, such as Dutch. In my eyes, a rigid set of definitions that have been proven to work in a rigid SVO word order language only are not the way to go.

The definitions given in the last few paragraphs provide a clear and solid starting point for this research, and they are compatible with these types of coordinated ellipsis in languages with a word order different from Dutch or English, which makes them applicable in a much wider environment than the old definitions I used as a starting point.

[8] This is a sentence from the complete Clef corpus, and was not included in the selection I studied. Courtesy of Gosse Bouma.

3. METHODS

To conduct the frequency analysis of ellipsis, I used a selection of two different Dutch corpora, one spoken and one written, so as to make a comparison with Meyer's study feasible. The spoken corpus consists of a 86,347-word selection (4875 sentences) from the Corpus Gesproken Nederlands (CGN for short), a collection of conversations ranging from interviews to soccer coverages. The written corpus contains 192,219 words (13448 sentences) from the Dutch Clef corpus, consisting of articles from the 1994 and 1995 editions of the Dutch newspapers Algemeen Dagblad and NRC Handelsblad[9]. Though Meyer used a corpus with more variation, my corpus is larger. There still are some differences though, apart from the language. I expected to find less right node raising, a form of ellipsis best suited for dry factual text, since the corpus I used, unlike Meyer's, doesn't contain legal text.

These subcorpora were first automatically parsed with Alpino annotate their syntactic structure. Since Alpino isn't able to correctly parse ellipsis yet, I had to find search patterns for Alpino that exclusively targeted all elliptical coordinations for right node raising, conjunction reduction, and gapping respectively.

I started with a simple search for coordinated sentences, which by our definition should encompass all sentences with the types of ellipsis this study focuses on. From there on I categorized the first 250 coordinated sentences from both corpora by hand, based on the presence or absence of ellipsis and the type of ellipsis if present. From this list I tried to determine the common characteristics of each type of ellipsis and condense these to a search pattern with which I could then scan the corpora for more ellipsis.

3-1. Alpino

According to its developers (Bouma, Van Noord, and Malouf, 2001), Alpino is a wide-coverage computational analyser of Dutch which aims at accurate, full, parsing of unrestricted text. Ideally this means that ellipsis should already be parsed correctly by the Alpino parser, however, this not to the case, and the possibility of finding ways to improve Alpino's parsing of ellipsis was one of the motivations for this study. Alpino is one of the best automatic parsers for Dutch available. Alpino's features an extensive grammar based on the OVIS grammar (van Noord, Bouma, Koeling, and Nederhof, 1999; Veldhuizen van Zanten, Bouma, Sima'an, van Noord, and Bonnema, 1999) which is in turn inspired by the Head-driven Phrase Structure Grammar (Pollard and Sag, 1994), and it includes rules covering the basic constructions of Dutch as well as more specific rules for individual cases.

The key elements that render Alpino-parsed sentences searchable for information are the dependency structures that the program automatically assigns to each sentence (Van der Laan, Bouma, Van Noord, and Malouf, 2002). Through these features one can deduce the grammatical relations that hold in and between the constituents of a sentence. Word order is also preserved, through labelling words with a "begin" tag and an "end" tag. Also, Alpino places indices to indicate a grammatical relation that carries over between two (or even multiple) different words. An example would be the sentence *"Hij kan een auto kopen." ("He can buy a car.")* in Figure 1 below[10]. Alpino indexes the subject "hij" with a number (1 in this case) and places an empty element referring to the subject in the subclause "een auto kopen" to indicate that "he" is in fact the subject linked to the verb "buy" (also note that word order is preserved in the tree, this is also the case with larger sentences). I used this example to show that indexing isn't solely used for ellipsis, but also for signalling inter-word relations in a sentence. Indexing is very useful, since ellipsis would be most accurately parsed with the

[9] For a complete listing of the used corpus parts, see Appendix B – Corpus Selections

[10] For the XML-trees in this paper I used the online parser. The reason is that it returns clear black-and-white trees that can easily be incorporated in this paper.

appropriate index at the point of the omitted word. For example, the sentence *"Jan drinkt koffie en Marie thee." ("Jan drinks coffee and Mary tea.")* would be indexed as *"Jan drinkt$_1$ koffie en Marie Ø$_1$ thee." ("Jan drinks$_1$ coffee and Mary Ø$_1$ tea.")*. Since an index carries the same grammatical tag as the word it refers to, the sentence can then be parsed normally, as though no words were elided.

Alpino, like other parsers, produces a multitude of dependency trees for each sentence. After the program is done producing these comes the task of selecting the best parse for each of those sentences. Usually, this is done automatically by Alpino, but it is also possible to do this by hand, with the help of several computational tools. It is also possible to correct the parses Alpino generates, but I purposefully used corpora not corrected this way, nor containing parses that were chosen by hand, as one of the aims was to see if Alpino itself parses ellipsis correctly.

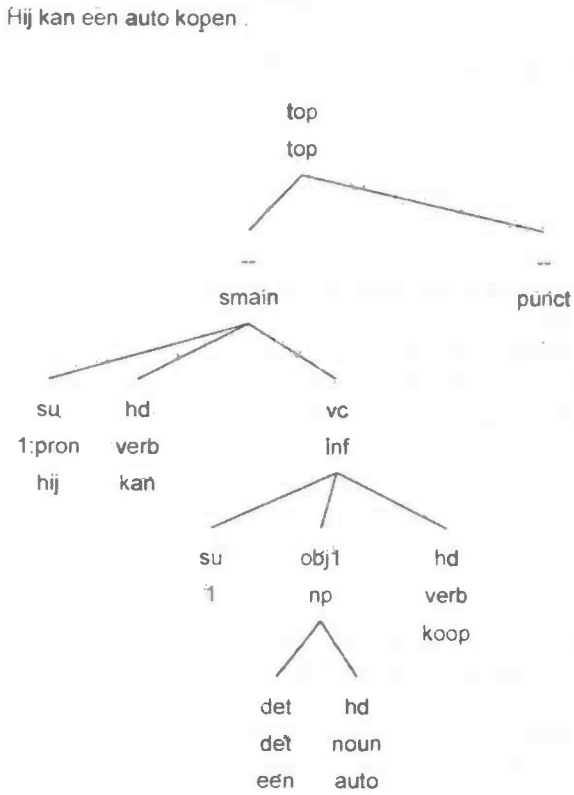

Figure 1.

3-2. XML-Querying

After all the words have been tagged and Alpino has chosen the parse it thinks is best, the resulting dependency tree is stored in the XML format (see Appendix C – Sample XML File). I used the XPath standard (ref: www.w3.org/TR/xpath) query language. A tool appropriately called dt_search enables me to use regular expressions to search the dependency trees for linguistic patterns. Chapter 5 discusses the final patterns in detail, but I will cover some important features here to facilitate understanding of those more complicated patterns later on.

```
//node[@cat="conj"]
```

This pattern searches for a node of the category "conj". The double slash at the beginning means that the place of that node in the tree is not important. Thus, any sentence with a "conj"

anywhere will match this pattern. Incidentally this search patters should find all ellipsis candidates, since ellipsis can only take place in a coordinated sentence. E.g. "Gisteren droegen [Jan en Joke] mooie kleren." ("Yesterday, [Jan and Joke] wore nice clothes.") has a conjunction (the word "en"/"and") in the middle of the sentence.

If I want to narrow down the search, daughters can be added. For example:

```
//node[@cat="conj" and node[@cat="np"]]
```

This will search for all sentences with a conjunct that features at least one "np" as a daughter. The sentence "[De boer en de slager] vieren feest." ("[The farmer and the butcher] party.") would be a good example.

XPath also enables me to use negation in my queries. Again, an example:

```
//node[@cat="conj" and not(node[@cat="np"])]
```

The pattern will now look for any sentence with a conjunct *not* featuring an "np" as a daughter. A match coverage for a game of soccer might produce such a sentence, like "[Piet en Kees] trappen af." ("[Piet and Kees] kick off."), since names are tagged as having pos (part of speech) "name" and no category[11].

Further constructions I can search for with the help of XPath are disjunction and comparison of numeric values. The last one is useful to make sure a word of a certain type is followed by a word of another type, since word order is encoded numerically with "begin" and "end" tags that indicate its place in a sentence. The CGN also has an @id tag that indicates the place of the tagged item in the parse tree, again encoded numerically.

[11] Names are not raised to NP-level by Alpino. Their relation to the other sentence elements (as subject, object, or otherwise) is tagged however in the @rel tag.

4. CORPUS STUDY

Each form of ellipsis corresponds to a unique syntactic pattern, and thus requires a separate search pattern. This is reflected in my corpus study, where I discuss the three types separately.

4-1. Conjunction Reduction (CR)

Conjunction reduction should, in theory, be easy to detect, as it leaves behind two constituents of the same type on each side of the conjunction: a verb and an object. The conjunct clause can thereafter be combined with the subject to form a grammatical sentence. Since a subject is a required part of any grammatical sentence, it is expected that Alpino will recognize the absence of the elided subject and parse such a sentence correctly. Even without the subject you're still left with perfect eligible Dutch.

In contrast to right node raising and gapping, conjunction reduction was very common in both the CGN and the Clef corpus. The CGN contained 28 reduced and 13 unreduced, and the Clef corpus 41 reduced and 1 unreduced instances of CR in their first 250 coordinated sentences. With unreduced I mean sentences where CR could have been used but wasn't. Most of the time Alpino got it right, and when it did go wrong it was because of lexicon problems. Example 30 shows a sentence that was parsed correctly by Alpino.

Example 30. (Clef Corpus)
[De pijl was 70 centimeter lang en Ø had een doorsnee van tien centimeter].
[The arrow was 70 centimetres long and Ø had a diameter of ten centimetres].
The arrow was 70 centimetres long and had a diameter of ten centimetres.

As you can see Alpino noticed that the subject was missing and inserted an empty element with index 1. The index must match with another node of the same relation, and thus the subject "de pijl" is co-indexed with number 1. So, even though the index has no special meaning, it does signal the absent subject. The tree in Figure 2 on the next page shows that the word group "de pijl" is "recognized", through co-indexing, as also being the subject of the following conjunct, and the sentence is parsed correctly. Alpino is very consistent in parsing conjunction reduction correctly as the following examples show.

Example 31. (Clef Corpus)
[De huidige Franse nummer één won in de verlenging van het Spaanse Joventut Badalona (90-86) en Ø streek naast de toernooiwinst ook nog eens een overwinningspremie van f 20.000 op].
[The current French number one won in the overtime from the Spanish Joventut Badalona (90-86) and Ø pocketed besides the tourney win also still once a winning prize of f 20,000.].
The current French number one beat Spanish Joventut Badalona in overtime (90-86) and pocketed not only the tourney win, but also f 20,000 in prize money.

Example 32. (CGN)
[Kinderopvang Nijmegen kijk*a krijgt straks vier lokalen en Ø wordt daarmee de grootste buitenschoolse opvang van Nederland].
*[Nursery Nijmegen look*a gets later four classrooms and Ø becomes therewith de biggest outschool nursery from the Netherlands].*
Nursery Nijmegen will be expanding to four rooms and thereafter be the largest out-of-school nursery in the Netherlands.

The * in the sentence above seems to denote immediately corrected mispronunciation and is apparently ignored in the parse.

De pijl was 70 centimeter lang en had een doorsnee van tien centimeter .

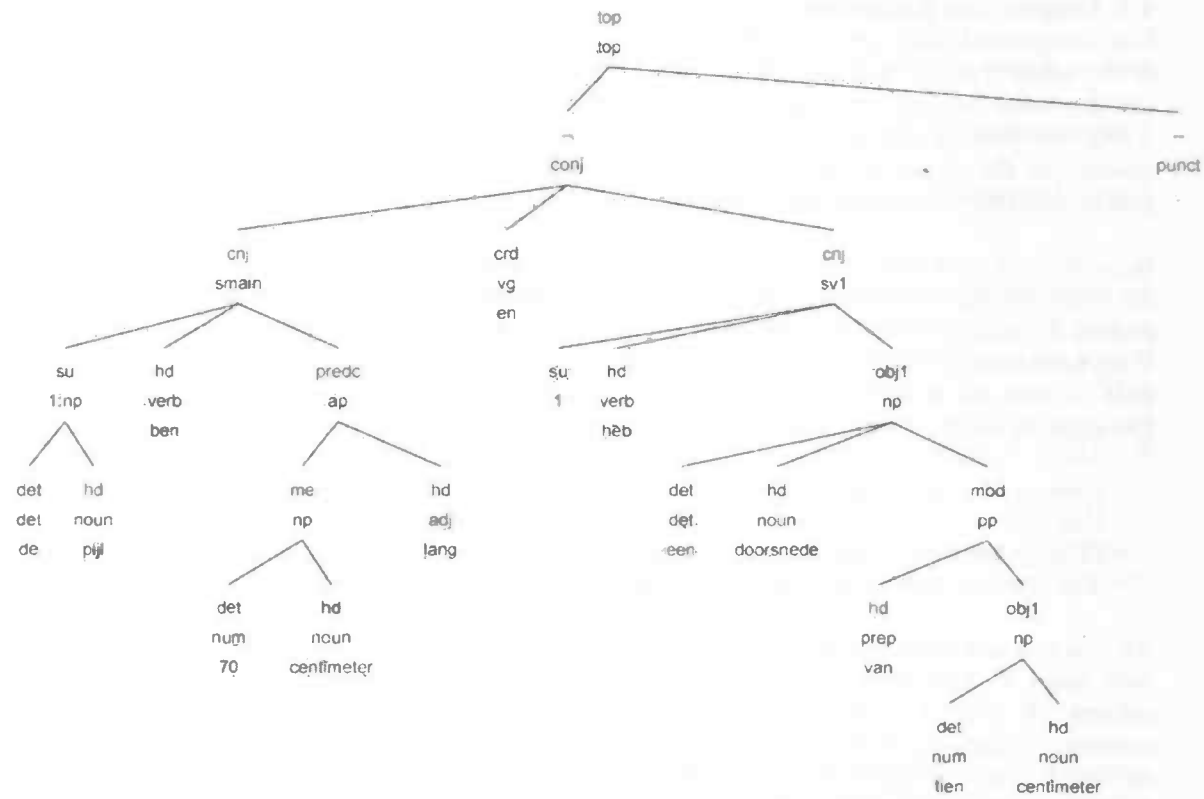

Figure 2.

Example 33. (CGN)
[bal wordt dus weggemaaid door IJsbrandy maar Ø is weer in Treffers-bezit].
[ball is being kicked away by IJsbrandy but Ø is again in Treffers-possession].
The ball is being kicked away by IJsbrandy but is once again in possession of the Treffers.

Alpino has major problems with ungrammatical input though, and if a sentence with grounding words (like *uh* for example) or missing punctuation is presented, the parser no longer seems so robust. Take for example the following sentence, taken from the CGN:

Example 34. (CGN)
bal wordt gepasst op Schaaij [Schaaij is heel druk voorin maar kan de bal niet goed meekrijgen].
ball is passed on Schaaij [Schaaij is very busy in the front but can the ball not good take with him].
ball is passed to Schaaij, Schaaij is very busy in the front but doesn't quite manage to take the ball with him.

At first sight there is nothing wrong with the text parse, Alpino selected the correct coordination after all. If we use the exact sentence, with the missing comma that is, as input for the online Alpino parser, the XML-tree in Figure 3 is the result. The tree doesn't match the parse in the CGN, and actually shows a wrong parse, where "druk" is read as a noun ("pressure" in English) and made the subject of "de bal niet goed meekrijgen" ("doesn't quite manage to take the ball with him"). Add in a comma though, and you'll get Figure 4.

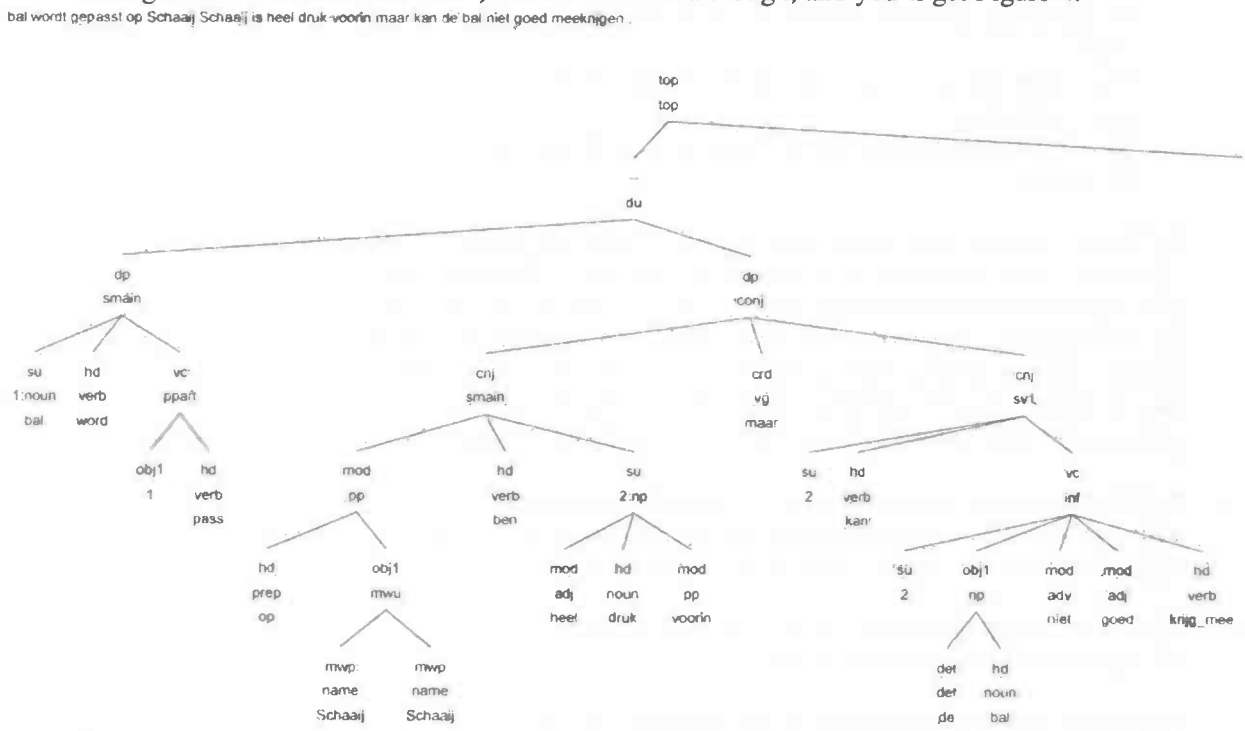

Figure 3.

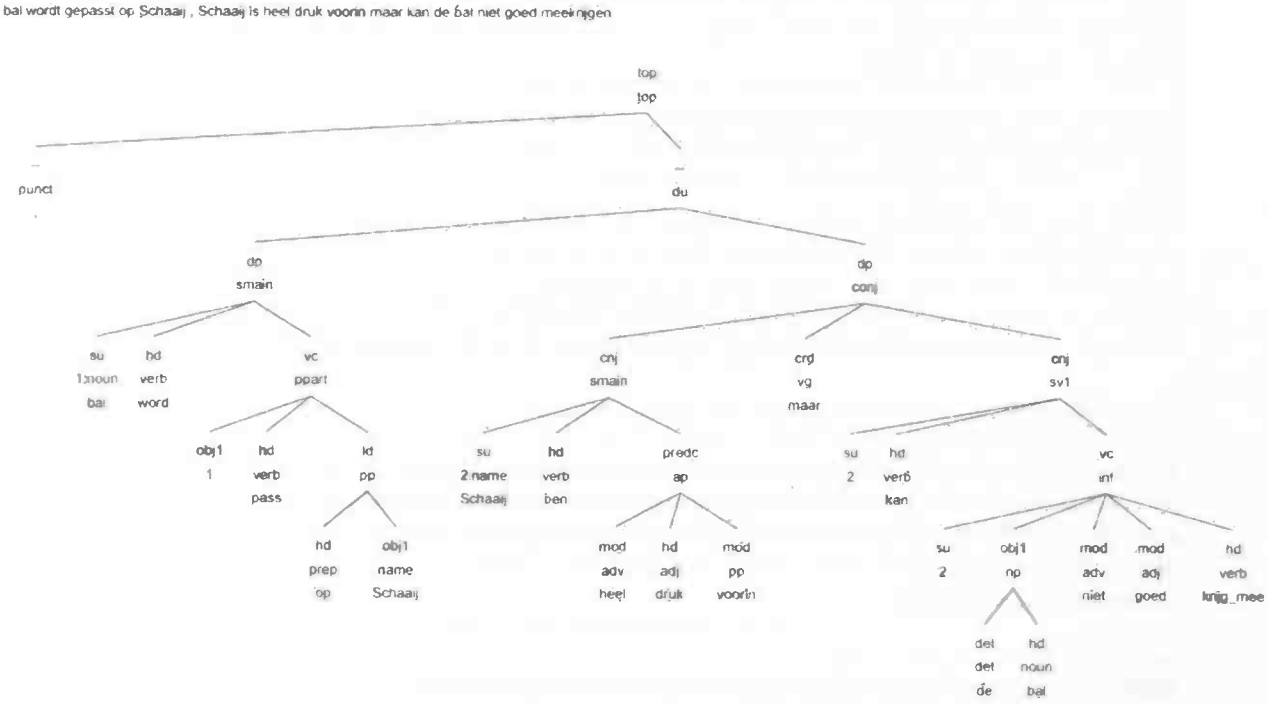

Figure 4.

Now this tree has a grammatically correct underlying structure that does lead to the text parse found in the CGN. There are a few more examples where bad punctuation or noise causes a wrong parse to happen, whereas the same sentence cleaned up would be parsed correctly. Let's take a look at the following sentence:

> ***Example 35. (Clef Corpus)***
> Er zijn altijd gevallen waarin ik me achter de oren [krab en me] afvraag wat ik er mee moet.
> *There are always cases wherein I me behind the ears [scratch and me] wonder what I there with must.*
> There are always cases where I scratch myself behind the ears and wonder what I should do with it.

In Dutch, *"krab"* can mean two things. It can be either the singular present tense of *"krabben"* (*"to scratch"*) or it can mean *"lobster"*. What happens in this particular case is that Alpino incorrectly assumes that *"krab"* is used as a noun, instead of inferring the use of the word group *"zich achter de oren krabben"* (*"to scratch oneself behind the ears"*). This enforces an impossible reading of the sentence and therefore leads to an incorrect parse. I see this as being a lexicon problem, where the wrong Part of Speech tag is assigned. Luckily, I came across only a few of these in the sub-corpora of conjunct sentences.

Problem sentences aside, there is a common characteristic of the correctly parsed sentences with conjunction reduction, namely that of a conjunctive followed by an indexed noun or NP. The following XML-node is typical for a correct parse.

```
<node rel="cnj" cat="sv1" begin="0" end="13">
    <node rel="su" index="1" />
```

This is the XML representation of the first word in a following conjunct, where a reference is made (through indexing) to a subject used earlier in the sentence. The following conjunct is classified as belonging to the "sv1" category, which means it is a verb sentence with a missing subject[12]. Another example.

```
<node rel="cnj" cat="ssub" begin="10" end="23">
    <node rel="su" index="1" />
```

In this case the following conjunct is labeled as a "ssub" category, which means that it's part of a subordinate clause. Both examples have at their top a node with a "cnj" relation – which is expected, since they are indeed part of a conjunct – and an indexed node with a "su" relation. A very good start for a search pattern would be to look for just that, a node with rel="cnj" with a daughter with rel="su" and an index. In XML-querying:

```
//node[@rel="cnj" and node[@rel="su" and @index]]
```

To check is the @index of the subject always equals "1" I tried the following sentence on the online parser, in order to check if Alpino will index the site of conjunction reduction with a 1, even if there is a previous site of ellipsis in the same sentence. If successful this will narrow down the search pattern even more, resulting in a better fit. Below the example is the corresponding parse tree.

---

[12] This is actually the incorrect category, since the indexed subject completes the sv1 to form a main clause of category "smain". This minor mistake does not influence my current search patterns, though it bars automatic rejection of sv1 clauses, which shouldn't normally feature a subject, indexed or not.

Example 36.
Jan koopt gebakken banaantjes en appeltjes en eet ze op.
Jan buys baked small bananas and small apples and eats them up.
Jan buys baked bananas and apples and eats them.

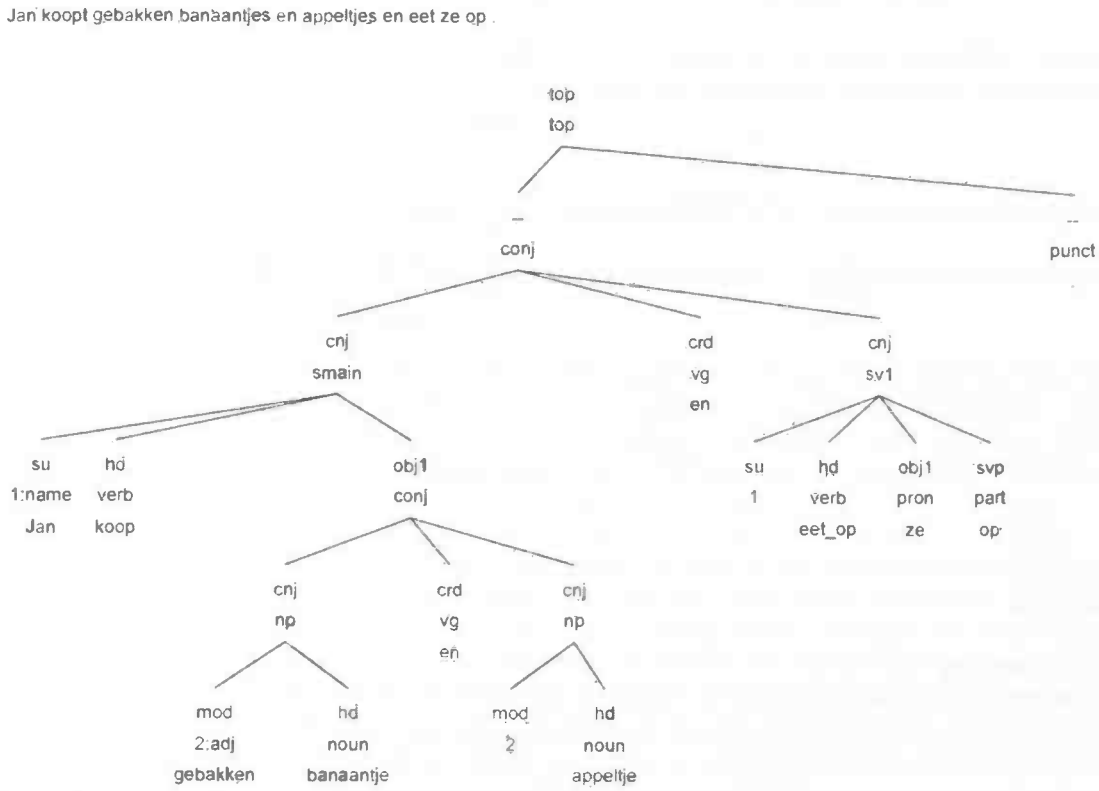

Figure 5.

Alpino notices the ellipsis, but still indexes the site of conjunction reduction with a 1. This is promising, because it might mean that we can add that to our search pattern. Still, the word group that is subject to conjunction reduction might contain ellipsis itself, so we have to check if that doesn't cause the index to shift to a higher number. I used the following sentence to test this.

Example 37.
De gebakken banaantjes en appeltjes zijn lekker en verkopen goed.
The baked small bananas and small apples are tasty and sell well.
The baked bananas and apples are tasty and sell well.

The tree on the following page shows that the complete subject conjunct is assigned an index of 1, and the elided words in the subject are assigned a higher index number. By now it is obvious that Alpino does not assign an index number based on the site of ellipsis, but rather based on the site of the antecedent. Based on this information I am going to try the following search pattern.

```
//node[@rel="cnj" and node[@rel="su" and @index="1"]]
```

This will probably still include a lot of rubbish, but we can refine the pattern later on if this doesn't work. Thus, with a good idea for a search pattern, it's time to look into gapping.

De gebakken banaantjes en appeltjes zijn lekker en worden goed verkocht .

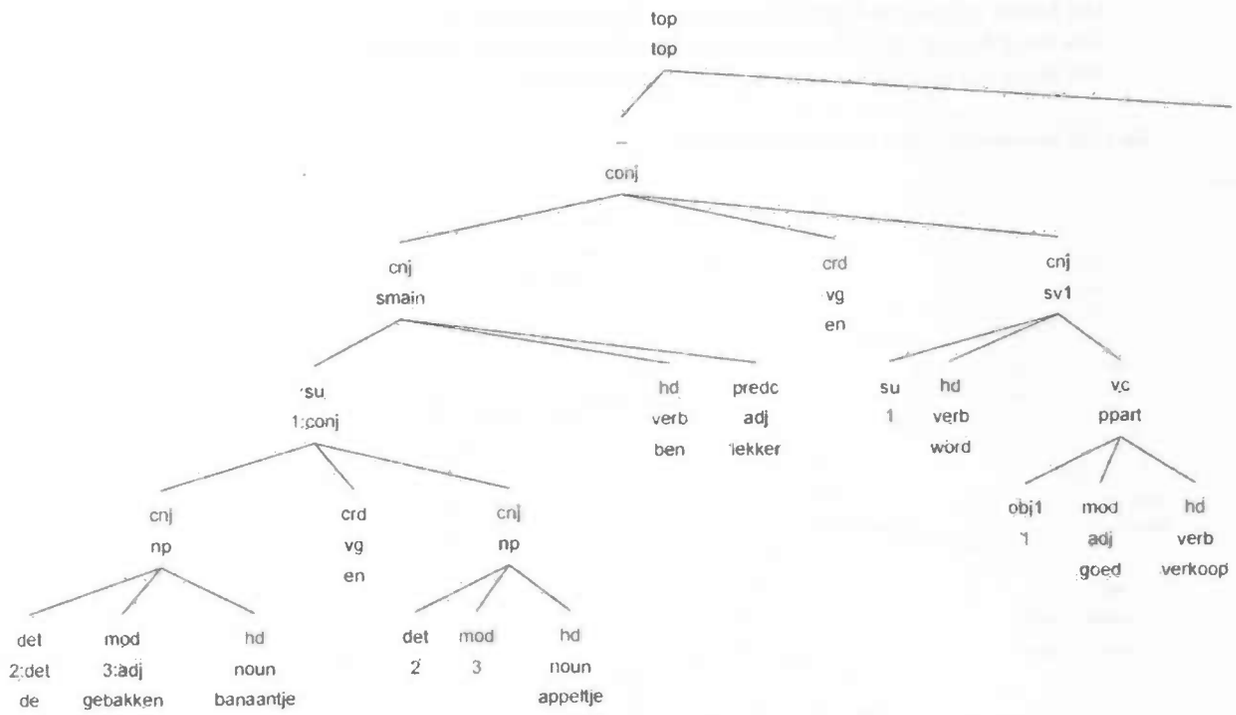

Figure 6.

4-2. Gapping

That gapping is a difficult form of ellipsis to parse for Alpino became clear during my corpus study. Unlike CR, gapping occurs in SVO word order sentences (which is the order of main clauses in Dutch) as often as in sentences with another surface structure, like SOV (subordinate clauses) or VSO (questions and PP-fronted sentences). In addition, in SVO word order, and often in VSO word order as well, gapping leaves behind constituents that don't match, as shown in the examples below. Regardless of word order though, Alpino was unable to produce any right parses on sentences with gapping. With 8 cases of gapping out of 250 coordinated sentences in the Clef Corpus and 1 out of 250 in the CGN, this kind of ellipsis might be just common enough to be of interest in further research though. Below are a few corpus examples of wrongly parsed gapping sentences.

Example 38. (Clef Corpus – SVO)
De omzet van bloemen steeg dit jaar met [4,2 procent en de prijzen] gemiddeld met 3 procent.
The turnover of flowers increased this year with [4.2 percent and the prices] on average with 3 percent.
The turnover of flowers increased with 4.2 percent this year, and the increase in price averaged 3 percent.

Example 39. (Clef Corpus – PP & VSO)
Voor HFC scoorden [Timmeren, Molenaar en Bruinink] ; voor de oud-internationals [Tol en Houtman].
For HFC scored [Timmeren, Molenaar and Bruinink] ; for the old-internationals [Tol and Houtman].
Timmeren, Molenaar and Bruinink scored for HFC ; Tol and Houtman for the old-internationals.

More examples could be given, but they all amount to one observation: Alpino is incapable of recognizing gapping. This might not be surprising, considering that it is hard to look past the messy syntax left behind in a gapped sentence. The usual trick of simple looking for a coordination of two word groups of the same category doesn't work here, because it is impossible to find a preceding and a following conjunct whose category matches.

To show what it should have been like, I give correct parses of the above sentences below. Remember, the square brackets indicate coordination, the underlined word is the antecedent which correlates with Ø, the empty element.

Example 40. (Clef Corpus)

a. [De omzet van bloemen steeg dit jaar met 4,2 procent en de prijzen Ø gemiddeld met 3 procent].

b. [Voor HFC scoorden [Timmeren, Molenaar en Bruinink] ; voor de oud-internationals Ø [Tol en Houtman]].

It seems that Alpino once again has trouble with the large scope in which ellipsis must be determined, preferring to take the shortcut, resulting, however, in limited coordination and thus an incorrect parse. Still, since we're looking for search patterns, maybe the parse trees will point out a common mistake in those sentences. I've cleaned up the input a bit to exclude mistakes due to noise.

As you can see, Alpino is unable to find a correct parse in both of the sentences. Furthermore, each one is parsed in a different way, because of a different word order, preventing me from formulating a single search pattern to at least find the wrong parses. In Figure 7a we see that "4,2 procent" is falsely coordinated with "de prijs". In Figure 7b, Alpino ignores the comma and reads "voor de oud-internationals Tol en Houtman" as a PP. My only chance of correctly identifying gapping therefore, is to construct patterns based on word order, since that is preserved in the xml-tree, even if Alpino fails to find a correct parse. For this, a search pattern which looks for a noun following a noun might work.

De omzet van bloemen steeg dit jaar met 4,2 procent en de prijzen gemiddeld met 3 procent .

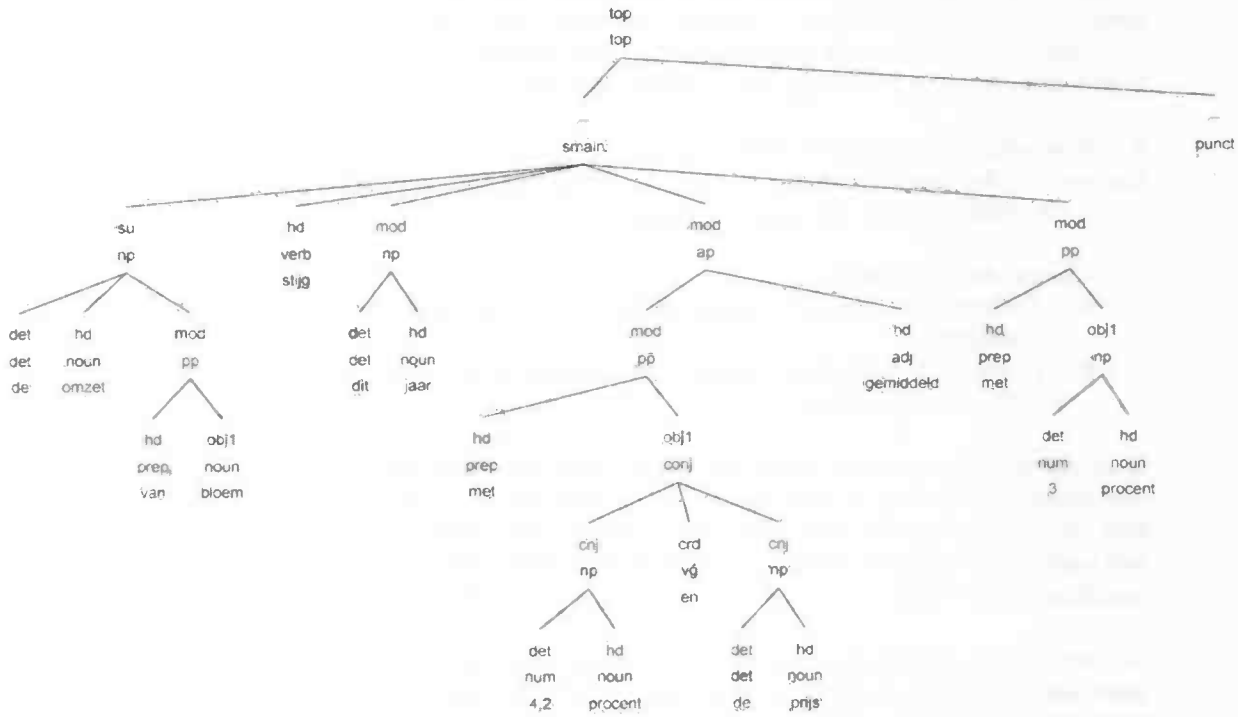

Figure 7a.

Voor HFC scoorden Timmeren , Molenaar en Bruinink , voor de oud-internationals Tol en Houtman .

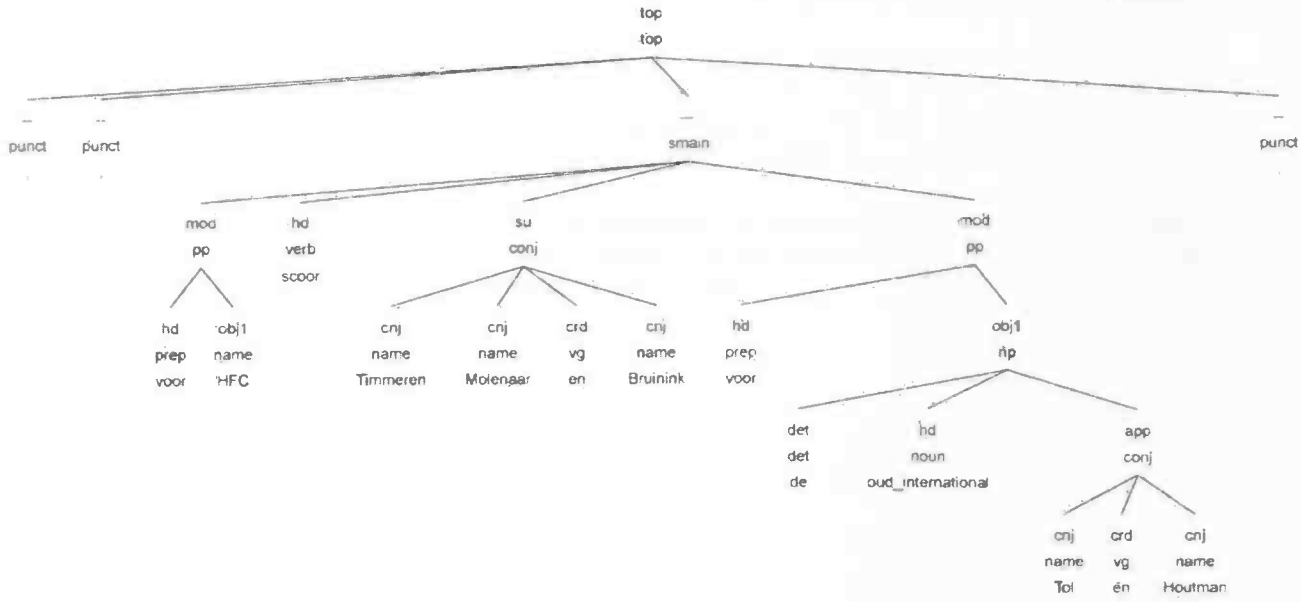

Figure 7b.

4-3. Right Node Raising (RNR)

Right node raising, the last and least common type of ellipsis, should not, in theory, form a problem for Alpino. Like CR, this type of ellipsis leaves behind two constituents of the same type on both sides of the conjunction, usually a subject and a verb. A grammatical parse is possible by simply coordinating these sides and then (since a coordination shares the category of its daughters) forming a sentence with the object. Sadly, in the 250 conjunct sentences of each corpus I found no cases of RNR.

To be able to make an accurate guess of how Alpino does parse RNR if it would come across it, I tested the online Alpino parser (http://ziu.let.rug.nl/vannoord_bin/alpino) with the following sentence.

Example 41.
Ik koop Ø en hij verkoopt brood.
I buy Ø and he sells bread.
I buy and he sells bread.

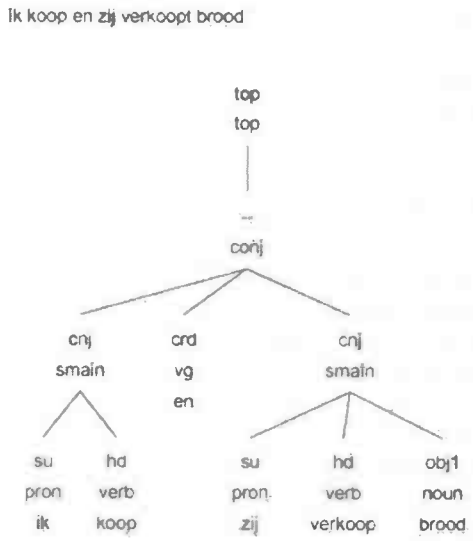

Figure 8.

At first sight this parse seems fine, but note that there is no indexing. This implies that Alpino simply parses *koop* as an intransitive verb. This would be a wrong parse after all because RNR is the elision of an object, and a sentence with an object requires a transitive verb. Suspecting that this might be the case I tried the following sentence, using the verb *plaatsen* to place), which can only be used transitively in Dutch.

Example 42.
Ik plaats Ø en hij doet een bod.
I place Ø and he does an offer.
I place and he makes an offer.

This led to the following tree.

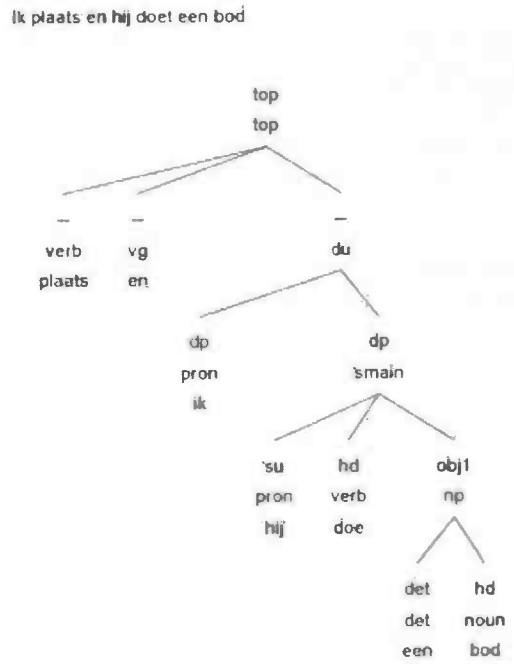

Figure 9.

When presented with a verb that *must* be transitive but where the object is elided, Alpino still can't find the correct parse. Moreover, Since it is impossible to construct a reading where *plaatsen* is intransitive, Alpino can't even produce a parse that looks right on the surface. Note that the search pattern can't be based on the wrong parse in Figure 9, since that would exclude cases of RNR like that in Figure 8. It seems that RNR forms a problem for Alpino after all, most likely because the rules for forming a sentence with a missing object don't exist.

Alternatively, it might be possible to once again conduct a search based on the order of the words. Since Alpino retains information on the position of a word in the sentence, a search pattern looking for a conjunct immediately preceded by a verb and followed by a subject should be able to find at least some of the sentences with RNR, if any exist in the corpus.

With the combined knowledge of the new definitions of ellipsis and the way in which ellipsis deals with ellipsis it's time to put that knowledge to the test in the next chapter and look at how the final patterns came to be and what they found.

5. RESULTS

Despite Alpino's inability to parse right node raising and gapping in a correct, or at least a uniform, way, I composed search patterns for each type of ellipsis. Of course, the CGN, being a corpus of spoken Dutch, was much harder to process than the Clef, because it is less grammatical. Hence, some differences in results are to be expected. In table 1 below I displayed the results of my search patterns applied to the manually analysed first 250 coordinated sentences from each corpus. The first two rows identify the corpus and the type of ellipsis. The row labelled # lists the total number of sentences with in the corpus with that type of ellipsis. The column "Correct" lists the number of sentences correctly labelled as the indicated type of ellipsis, the column "False Alarm" lists the number of lines falsely labelled as that type, and the column "Missed" finally reports the number of sentences with that type that were not detected by the search pattern. So, for example, out of the first 250 in the Clef corpus, 43 coordinations contain Conjunction Reduction, of which 32 were recognized by the search pattern and 11 were not. The last two columns list the Recall and Precision scores of the patterns used. The following formulas were used to calculate these percentages.

Recall = Correct / (Correct + Missed)
Precision = Correct / (Correct + False Alarm)

In other words, Recall measures the fraction of relevant material that is returned by the search pattern relative to the number of actual cases of ellipsis in the entire set of 250 sentences, and Precision the percentage of relevant hits produced by that pattern among all sentences returned by the pattern.

		#	Correct	False Alarm	Missed	Recall	Precision
CGN	RNR	-	-	9	-	-	-
	CR	28	14	1	15	48%	93%
	Gap	1	-	13	1	0%	0%
Clef	RNR	-	-	0	-	-	-
	CR	43	32	0	11	73%	100%
	Gap	7	5	19	2	71%	21%

Table 1.

The sections below discuss these results for each of the three ellipsis types.

5-1. Conjunction Reduction

Because of the prevalence of grammatically correct sentences in the Clef corpus (as opposed to the large number of ungrammatical ones in the CGN) I started with Conjunction Reduction. My manual research showed that approximately 20% of all coordinated sentences may contain conjunction reduction (43 out of 250). The first search pattern I proposed on basis of the manual research in chapter 4 was the following.

Pattern 2. (Clef Corpus)

```
//node[@rel="cnj" and node[@rel="su" and @index="1"]]
```

As hoped for, this search pattern found all occurrences of normal Conjunction Reduction (i.e. CR in SVO word order sentences) I found by hand in my manual study of the first 250 coordinated sentences. Unfortunately, the pattern also returned many sentences without conjunction reduction, the search pattern clearly needed refining. To get a clear idea of what

goes wrong with this search pattern, take a look at the following two trees. The two nodes that correspond to the pattern are boxed. Both sentences are taken from the Clef Corpus (once again, the part of the tree representing punctuation is cut of).

De pijl was 70 centimeter lang en had een doorsnee van tien centimeter .

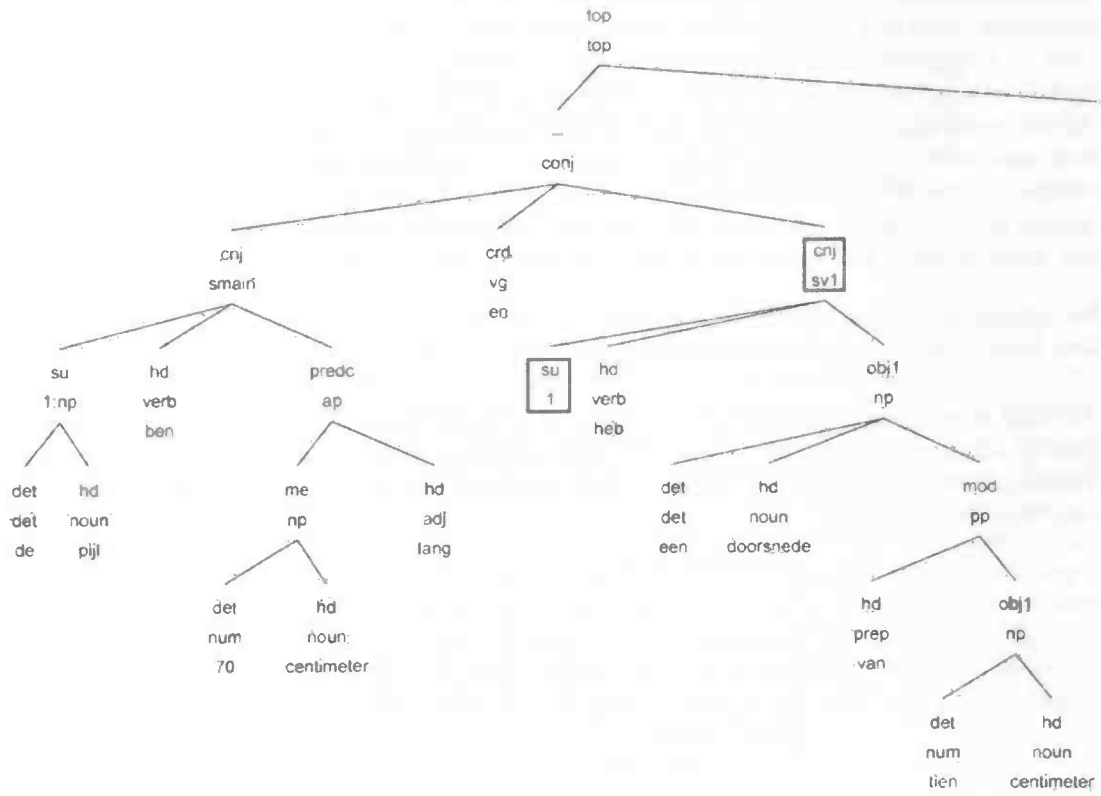

Figure 10a.

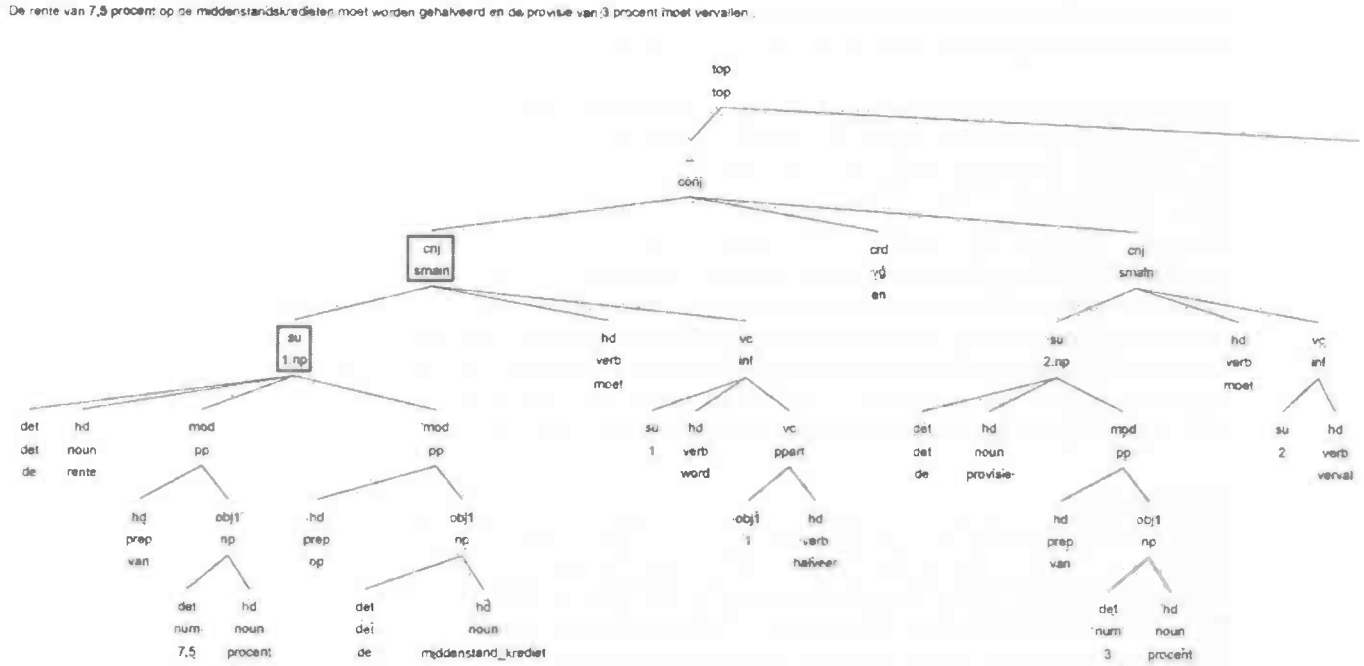

Figure 10b

If we look at the underlying XML-tree, we can see the difference between those two sentences however. The first piece of XML-code below belongs to the tree in figure 10a, that is, the sentence which does feature conjunction reduction. The second belongs to the sentence in figure 10b, which contains no ellipsis.

```
<node begin="0" cat="sv1" end="13" id="13" rel="cnj">
    <node begin="0" end="2" id="14" index="1" rel="su" />

<node begin="0" cat="smain" end="11" id="2" rel="cnj">
    <node begin="0" cat="np" end="8" id="3" index="1" rel="su">
```

We've already seen that the category of the parent node in ellipsis may vary, the main difference therefore is that the sentences (see also chapter 4-1) with conjunction reduction have no category attached to the indexed site of ellipsis, whereas the false alarm does have a category, namely "np". One other difference is of course that the real example of CR has its site of ellipsis directly after the conjunct *en*, which has an @cat="crd". Other than that however, the two XML-trees are eerily similar, and if changes to the search pattern incorporating these differences won't set the two apart, it's going to be tough to find a working pattern to find only conjunction reduction.

Pattern 1. (Clef Corpus)

```
./dt_search_clef -s '//node[@rel="cnj" and node[@rel="su" and @index="1" and
not(@cat)] and node[@pos="verb"]/@begin=//node[@rel="crd"]/@end]'
```

The pattern above searches for a coordinate sentence with an indexed subject without a category that is immediately preceded by a conjunct. When I tested this pattern on the Clef corpus and compared the results with the part of the corpus I research by hand, it yielded a recall score of 73% and a precision of 100%. It returns only sentences with an SVO word order, but it proves that those are parsed correctly by Alpino (apart from the @cat, which

should be smain, since the indexed subject complements the verb and the optional object to form a main clause).

When tested on the part of the Clef corpus that wasn't manually tagged, very good results were yielded as well. A total of 356 sentences were returned by the pattern, of which I calculated the precision of the first 50 sentences not included in the manually tagged part. Of these 50 sentences only 3 were false alarms, corresponding to a 94% precision score. Strangely enough, the spoken CGN didn't return such good results. Only 17 additional sentences were found, and of those 7 were false alarms. This results in a precision of only 59%, much lower than the test set. I tried to determine exactly what happened here, and it seems like there simply is less SVO conjunction reduction in the rest of the corpus. Another explanation could be that the percentage of sentences where CR was possible but not used is higher in the rest of the corpus, resulting in fewer hits. All in all though, this pattern works well, especially in the (grammatically correct) written Clef corpus.

5-2. Gapping

As it turned out, gapping was the most difficult ellipsis type of the three (even though I did find some). Table 1 shows that for the Clef corpus the pattern I found has a recall of 63% and a precision of only 21%. The precision gets even worse in the rest of the corpus, showing 2% for the CGN and 4% for the Clef corpus. The main reason for the low precision is the fact that Alpino is unable to parse gapping correctly and turns it into an XML-tree that doesn't distinguish itself enough from other sentences. Additional problems arise because, though gapping is relatively common, it is distributed evenly over word order in Dutch. That is, gapping is as common in main clauses as it is in PP-fronted sentences and subordinate clauses[13]. Because I am unable to find a working pattern (or working patterns) for gapping I am unable to say much about it. Needless to say, this is a bit of a letdown, since it also prevents me from doing any frequency analysis and comparing the data with Meyer's research. The best pattern I could find is found below, split up in readable parts. The pattern is preceded by a small piece of code that tells dt_search to look in the Clef corpus, not shown below.

Pattern 2. (Clef Corpus)

```
a) //node[@rel="crd"]/@end=//node[@rel="cnj" and not(@word) and not(node[@pos="verb"])]/@begin
b) //node[@rel="cnj" and not(@word) and not(node[@pos="verb"])]/@begin=//node[@rel="det" or
      @pos="name"]/@begin
c) //node[@rel="cnj" and (node[@pos="noun" or @pos="name"]/@end=//node[@pos="adj" or @pos="det"
      or @rel="mod"]/@begin)]
d) //node[@rel="cnj" and not(node[@pos="verb"])]
e) //node[@pos="verb"]
```

It's a complicated compound pattern, but by splitting it up it becomes easier to understand and explain. The first part searches for a coordinate (@rel="crd"), immediately followed by a conjunct consisting of multiple words (hence the not(@word), but not a verb. The second part requires that that conjunct starts with either a determiner or a name[14]. The determiner forms a subject together with a noun, and the name is a subject in and of itself. Line c) states that there must be a noun or a name followed by a determiner or an adjective or a modifier, and basically searches for a subject and an object, the second immediately following the first. The fourth line then makes sure there is a conjunct without a verb in the whole sentence. Finally,

[13] Note however that main clauses are more common than PP-fronted sentences, meaning that the net number of gapping cases in main clauses will be higher than the number of PP-fronted gapping cases.

[14] A conjunct can also start with a noun, as in "Cats love fish and dogs bones", but adding @noun drastically lowered the precision and didn't increase recall.

part e) requires that there must be a verb somewhere in the sentence, which filters out short sentences, mostly answers to questions and the like, e.g. "With milk and sugar please."

In short, it searches for a conjunct without a verb and a subject followed by an object. Because the sentences are not parsed correctly, it is impossible though to search directly for nodes labelled as such.

5-3. Right Node Raising

As shown by table 1, neither the CGN nor the Clef corpus contains Right Node Raising in the first 250 coordinated sentences. I composed a pattern nonetheless, to look for cases of RNR in the rest of the corpora. The pattern is based on a parse made with the online Alpino demo (http://ziu.let.rug.nl/vannoord_bin/alpino). Consider briefly the following example.

Example 43a. (transitive)
Jan koopt Ø en Piet verkoopt auto's.
Jan buys Ø and Piet sells cars.
Jan buys and Piet sells cars.

This is how a correct parse of RNR would look. However, *kopen* isn't restricted to a transitive reading. Basically, I use the observation that Alpino chooses the path of the least resistance, which means that RNR is parsed as a normal sentence as long as the verb in the preceding conjunct can be intransitive as well. In this case the order of the constituents will consist of a verb, followed by a conjunct, followed in turn by a non-indexed subject as in the example below.

Example 44b. (intransitive)
Jan koopt en Piet verkoopt auto's.
Jan buys and Piet sells cars.

The patterns used vary because of the @id element present in the CGN, but not in the Clef corpus, as discussed in chapter 3 – methods.

Pattern 3a. (CGN)
```
`(//node[@pos="verb"]/@id + 1)=//node[@rel="crd"]/@id and
(//node[@rel="crd"]/@id + 2)= //node[@rel="su" and not(@index)]/@id'
```

Pattern 3b. (Clef Corpus)
```
`//node[@pos="verb"]/@end=//node[@rel="crd"]/@begin and
//node[@rel="crd"]/@end=//node[@rel="su" and not(@index)]/@begin and
//node[@rel="cnj" and node[@rel="obj1"]] and //node[@rel="cnj" and
not(node[@rel="obj1"])]'
```

Both share the same base pattern to search for a verb followed by a conjunct, and a conjunct followed by a non-indexed subject. The second part of the Clef pattern asks for a sentence with both a conjunct containing an object, and a conjunct without an object (the order in which these appear in the sentence is ignored). I left this part out of the CGN pattern because that resulted in 0 matching substrings. The number of sentences returned with the more general pattern was small enough to check by hand, so I chose the less specific pattern to be able to better determine if RNR occurred at all in the corpus. After being applied to the entire corpus, both those sentences returned by the Clef pattern and those returned by the CGN pattern resulted in no further findings of RNR. Additional patterns that looked for RNR in VSO (example 45a below) and modified SVO (example 45b below) word order sentences were also unable to find cases of RNR. Notice that 45b is in fact a fairly rare kind of sentence,

as it needs the object to be a unique item (*the* book as opposed to *a* book) to really work. In light of these tests I think it likely that there are no cases of RNR present in these two corpora.

Example 45a.
Gisteren kocht Jan Ø en verkocht Piet <u>een auto</u>.
Yesterday bought Jan Ø and sold Piet <u>a car</u>.
Yesterday, Jan bought and Piet sold a car.

Example 45b.
Jan leest geroerd Ø en Piet leest onverschillig het <u>boek</u>.
Jan reads touched Ø and Piet reads indifferent the book.
Jan reads the book emotionally and Piet reads the book indifferently.

In a final attempt to find RNR in the corpus, I used a different approach that builds upon the fact that Alpino incorrectly parses verbs in an RNR sentence as intransitive if there object has been elided. Figure 8 (page 33) shows a tree where the verb *kopen* (*to buy*) has been incorrectly parsed as intransitive. This fact can be used to search for intransitive verbs that are predominantly used transitively. Compiling such a list is easy enough, one only has to search through a corpus and count the number of times a particular verb is used transitive and the number of times it is used intransitive. Then, if a verb is encountered intransitive, but normally used transitive, it is highly probable that RNR is involved. I wrote a Perl script (Appendix C – tally_verbs.pl) to tally the (in)transitivity of each unique verb in the corpus selections. All verbs that were used transitive more than three times as often as intransitive were listed. Those verbs used less than two times in a transitive way were weeded out to make sure the verbs on the list were well-used verbs as well as to make the list much shorter. The sentences in which the verbs on the final list were used intransitive were then checked manually for RNR. Though this manual researched showed that these predominantly transitive verbs were incorrectly identified as intransitive most of the time, unfortunately not a single case of RNR could be found with this method either, leaving me with the conclusion that RNR is a rarely used type of coordinated ellipsis indeed.

6. DISCUSSION

The previous chapter shows clearly that, while CR is handled adequately, Alpino is still a long way off from dealing with ellipsis. This section discusses the results in chapter five and also tries to find remedies for the problems found. In addition, this chapter goes into greater detail on some of the issues in the field of ellipsis and in Dutch ellipsis in particular that I havened addressed in this paper yet.

6-1. Spoken and Written Corpora

When the results of the manual study on the CGN and the Clef corpus are compared with each other, there are a number of things that strike me as important. First, there are 43 instances of ellipsis in the written corpus, and only 28 in the spoken corpus. That is, ellipsis is more than 1.5 times as common in the written Clef corpus as it is in the spoken CGN corpus. This is consistent with Meyer's findings, though the differences in his corpus were even greater[15]. Table 1 below shows his results, where *total* lists the total number of cases where ellipsis was possible. *Full ellipsis* lists the number of cases where ellipsis was used, *partial ellipsis* the number of cases where the antecedent was replaced by a pronoun, and *full form* the number of cases where ellipsis could have been used, but wasn't.

	Full Ellipsis	*Partial Ellipsis*	*Full Form*	*Total*
Speech	23 (37%)	2 (3%)	38 (60%)	63 (100%)
Writing	351 (66%)	39 (7%)	142 (27%)	532 (100%)

Table 1 – Meyer's Corpus Results

One explanation Meyer had for the difference in percentage of ellipsis between speech and writing was that semantically less dense discourse, i.e. unabridged speech, is easier to understand and therefore conveys the message better than speech containing ellipsis. It is easy, however, to add meaning to spoken conversation through non-linguistic means. One would suspect that knowing who you're speaking to and being able to use body language to get across your point provide enough context to enable ellipsis on a larger scale than in written text, where these things are just not possible (although you can of course write with an audience in mind). It appears however that this results in "ellipsis" of another kind, namely omitting common information in general. When given the possibility to use ellipsis in a grammatically correct sentence, often the speaker would just utter the uncondensed sentences.

Two possible contributors to explain the difference in prevalence of ellipsis in speech and writing that haven't been mentioned yet are the roles of both the writer and the reader in written text. The writer has more time to write down what he wants to say than a speaker has time to plan a sentence because the text can be submitted at a later time. Thus, the writer gets ample time to think over and plan the way he wants to put his text on paper, giving him the opportunity to add "time-consuming" elements like ellipsis. On the other side of the equation stands the reader, who has an easier time understanding ellipsis in written text, since he can simply reread the line if he doesn't get it the first time. This is something the writer can (and apparently does) take into account when writing. In effect, writing lowers the importance of the natural flow of speech argument and the semantically less dense discourse argument I talked about in the introduction of this paper, since memory issues aren't really an issue here. This intuitive line of thought does explain why ellipsis is more common in writing than in speech, which leads me to believe that the rule of Speaker's Economy of Effort might not be as important to ellipsis as initially believed. Moreover, because it doesn't simply compete with the natural flow of speech and semantically less dense discourse argument, but is

[15] Note that the spoken part of the corpus Meyer uses is much smaller than the written part. 16.000 versus 80.000 words. The percentages in table 1 give a better idea of the relative frequency of ellipsis in English though.

downright at odds with them, I believe another unknown driving force is behind the realization of ellipsis. Possibly, a rule of Speaker's Economy of *Output* is at work, pushing the speaker to produce as short a surface form as possible. A rule like this *can* compete with the aforementioned rules to create the tension seen in ellipsis, and thus explain the difference in realization of ellipsis between speech and writing.

Next, fully one third of the sentences in the CGN which could have been subject to ellipsis were not, as is shown in table 2a below. On the left side of the table the number of elided sentences is listed, and on the right the number of sentences where ellipsis was possible but not used. Since I didn't find any cases of RNR in the whole corpus, it was left out of the table. Note that all but one of the sentences where CR was not used are in an SVO word order. With such a small sample, it might mean nothing, but it would be interesting to see if this trend holds in further research. If it does, an explanation needs to be found as to why ellipsis is ignored in SVO clauses more often than in any other word order. In any case, the table shows, at least for CR, that the smaller number of elided sentences in the spoken CGN is not a result of fewer opportunities but really a result of the speaker opting to use ellipsis less frequently. If we calculate the numbers, we see that in 94% of the sentences where ellipsis was possible it was actually used in writing, and in 67% of the time in speech. Though these percentages are much higher than the 73% Meyer found for writing and 40% he found for speech, the trend is the same. In writing, ellipsis is much more common than in speech, not because the possibility of applying ellipsis is more frequent in writing but because that possibility is more frequently used. This means the results of my (manual) research support Meyer's findings and conclusions, in particular the influence of the natural flow of speech and the semantically less dense discourse arguments, since they are equally adept at explaining the distribution of ellipsis in Dutch as in English.

CGN	**CR**	SVO	20	**CGN**	**CR**	SVO	13
		SOV	5			SOV	1
		VSO	3			VSO	-
	Gap	SVO	-		**Gap**	SVO	-
		SOV	1			SOV	-
		OVS	-			OVS	-
		VSO	-			VSO	-

Table 2a: Ellipsis in the CGN

Clef	**CR**	SVO	37	**Clef**	**CR**	SVO	1
		SOV	5			SOV	-
		VSO	1			VSO	1
	Gap	SVO	2		**Gap**	SVO	1
		SOV	2			SOV	-
		OVS	1			OVS	-
		VSO	2			VSO	-

Table 2b: Ellipsis in the Clef Corpus

We can also see that gapping is a lot less frequent that CR. This means patterns for finding gapping in a corpus must have a much higher precision than those for CR, because we don't want a marginal improvement in correctly parsing a rather minor phenomenon at the cost of a higher imprecision in other areas.

A third interesting point is the distribution of ellipsis amongst sentences with different word order. The difference between CR and gapping is striking. In the CGN as well as the Clef corpus, around 80% of CR (including those cases where CR could have been used but was not) can be found in main clauses, whereas the distribution of gapping in the Clef corpus is evenly spread out among the four different word orders it was found in.

One explanation would be that retrieval of an elided part would be optimal if the antecedent is fronted, as in the example below. This explanation is supported by the serial position effect I explained in chapter 2, which states that items at the start (or end) of a list are more easily remembered than those in the middle.

Example 46.
Een aantal fans kon het niet meer verdragen en Ø bestormde het veld.
A number of fans could it not more bear and Ø and charged the field.
A number of fans couldn't bear it anymore and charged the field.

The subject and antecedent *een aantal fans* heads the sentence. Knowing that SVO word order is the surface structure of main clauses like example 46 in Dutch, it is only logical that most cases of CR occur in that form. This line of thought also predicts that CR would be more common in SOV word order than in VSO word order, which, by a small margin, it indeed is.

Gapping on the other hands revolves around the ellipsis of a verb, which means that a VSO word order will be favoured according to this explanation. Next in line is SOV, which places the verb at the end, and SVO is actually the least favoured for gapping. We know however that most sentences have an SVO word order, followed by SOV and VSO word order. The higher number of SVO and SOV sentences will therefore provide some possibilities for gapping as well, resulting in a seemingly even distribution.

Another explanation would be that, because in cases of CR the focus often lies on the subject, main clause word order is preferred, since fronting of the subject in part signals its focus. Verbs however are rarely the focus of a sentence, and thus word order isn't nearly as important, except maybe for VSO word order if the verb needs to be stressed. This lack of focus results in a more even distribution. The two examples of VSO ellipsis I found in the corpus seem to support this theory. In 47a, the focus lies on *scoorden*, which is the point of soccer after all, in 47b the focus lies on *aan het woord is*.

Example 47a. (CGN)
Voor HFC scoorden Timmeren, Molenaar en Bruinink; voor de oud-internationals Ø Tol en Houtman.
For HFC scored Timmeren, Molenaar and Bruinink; for the old-internationals Ø Tol and Houtman.
Timmeren, Molenaar, and Bruinink scored for HFC; Tol and Houtman for the former internationals.

Example 47b. (CGN)
Aan het woord is niet de Noordierse dominee Paisley, maar Ø de Nederlander ds G.H. Kersten.
On the word is not the North-Irish minister Paisley, but Ø the Dutchman minister G.H Kersten.
Not the North Irish minister Paisley is speaking, but the Dutch minister G.H. Kersten.

6-2. Detecting Ellipsis
If anything, this research has shown that ellipsis is difficult to find automatically. There are a number of ways to find ellipsis, some of which might be implemented in Alpino. First, in certain cases, semantic clues could point to the presence or absence of ellipsis. For example, if parsers could be stocked with a database of tagged verbs, divided into groups based on the type of subject and object they need, it would be possible to deduce ellipsis (in this case RNR) from the fact that two verbs of the same group are encountered in a coordinated sentence, but they don't follow a parallel structure. The example below shows what I mean.

Example 48a.
Jan koopt Ø en Piet verkoopt <u>een auto</u>.
Jan buys Ø and Piet sells <u>a car</u>.
Jan buys and Piet sells a car.

Example 48b.
Jan wast en Piet feliciteert Peter.
Jan washes and Piet congratulates Peter.
Jan washes and Piet congratulates Peter.

In the first example, to buy and to sell are both verbs that have the same general actors. The subject is generally a person, and the object can be anything that can be owned, including things like freedom or happiness. Example 48b shows an example of two verbs that get in the way of this type of elliptical reading. Though syntactically there is nothing wrong with the sentence, to wash and to congratulate don't have a common target. To wash can target anything physical, from a person to a house. To congratulate usually targets only persons however, resulting in a clash of interpretation. One last example shows that it need not be the verb that makes this distinction, it can also be the object.

Example 48c
Jan verft en Piet maait het weiland.
Jan paints and Piet mows the meadow.
Jan paints and Piet mows the meadow.

Here we have the meadow, which can be mown, but, being organic, is not usually the subject of a paint job. Of course it is impossible to use information like this with Alpino, but future automatic parsers could benefit from using an extensive database (certainly more elaborate than wordnet (http://wordnet.princeton.edu/)) like this when looking for ellipsis.

Secondly, gapping and right node raising each have characteristics that might be sought out in a text. Of course both occur in coordinated sentences, so that will be the first clue. Then, for gapping, if one of the conjuncts has a subject immediately followed by an object, if the subject in the other conjunct is followed by a verb, it is probable that we have a case of gapping[16]. Also, a conjunct with a subject and an object, but not a verb, is probably gapping as well.

For RNR one simple rule might make a lot of difference. If, in a coordinated sentence, both the preceding and the following conjunct feature a verb which can be transitive, but only the following conjunct has an object, than it is possible that the coordination features RNR. Even better, automatic parsers could make use of a simple list of verbs that are used transitive most of the time, like the one I used to try and find RNR after the initial pattern search failed. A prerequisite is that the automatic parser must correctly[17] identify the (in)transitivity of verbs in regular cases, but if it can, this is one of the most solid ways to detect RNR.

Following these two guidelines might enable Alpino to better detect Ellipsis in the future, despite the obvious limitations of its syntactic approach.

[16] Not that this check doesn't work in SOV or VSO word order sentences.
[17] Actually, if the preceding conjunct is subject to RNR, the verb should be "correctly" labeled as being transitive, even though the object is missing, but for the sake of the argument I assume that it will be initially labeled as intransitive, as long as RNR has not been detected yet.

6-3. Different Types of Ellipsis

McShane (2005) takes the definition of ellipsis one step further than I did by defining how the meaning of a missing element can be reconstructed: either from the context or from a person's knowledge of the world. This approach hints at two different categories of ellipsis, and McShane makes a convincing point. She distinguishes between semantic ellipsis, which depends on the hearer's knowledge of the world, and syntactic ellipsis, which is retrievable from context[18]. An example of the first, taken from her book, is

> ***Example 49. (McShane)***
> He is reading (a book written by) Tolstoy.

Some might argue that this is not ellipsis, but rather a kind of metaphor. The problem is that this doesn't explain why the above sentence isn't possible in Chinese, since we know metaphors in general are possible in Chinese (for example, Ning Yu, 2003). If we regard the *read an author* construction as a case of ellipsis however, we can explain this in part. That is to say, Chinese belongs to the most restrictive class of languages when it comes to allowing ellipsis (Sanders, 1977, in Eckman), which is why it should come as no surprise that precisely Chinese should block it in this case.

Another example:

> ***Example 50.***
> Blueberry pie is a favorite dish.

Everybody will assume that blueberry pie is a favorite dish for a group much narrower than what would be possible based on a syntactic reading of the sentence alone. Based on the knowledge of the interpreter and the context this sentence could read as *"Blueberry pie is a favorite dish (at the local tavern)"*, or *"Blueberry pie is a favorite dish (of grandmothers around the world)"*. Further evidence that knowledge of the world is used to interpret this sentence is the fact that no one will read it as meaning *"Blueberry pie is a favorite dish (of penguins)"*. This may seem obvious, but it implies we add information to be able to make sense of the sentence, which implies ellipsis on the semantic level.

An altogether different case of semantic ellipsis is shown below:

> ***Example 51.***
> Sue likes Beethoven.

The person in question doesn't literally like Beethoven, rather she is very fond of the classical masterpieces composed by him (either that or a cute 1992 movie starring a likeable dog with the same name). Either way, very few will read the sentence and settle for the literal meaning of Sue liking the person Beethoven.

In the end my research focuses on computational research though, and there is one problem with semantic ellipsis that can't be solved; it's reconstruction depends on assigning a meaning to the elided sentence based on knowledge of the world, and that is something automatic parsers like Alpino simply can't cope with, at least not until it becomes possible to completely model the world in a way that provides quick and easy access to the parser.

---

[18] Note that even syntactic ellipsis needs knowledge of the world to some degree. A context-based parser like Alpino can't tell the difference between "I drank and Pete bought a shovel" and "I drank and Pete bought apple juice", where the first sentence does not, but the second does contain RNR and knowing that you can't "drink a shovel" is essential for a correct parse.

6-4. Future Research

Much is left to be done after my research, in a number of different directions. Of course Dutch still needs good frequency data on ellipsis. Since it seems like automatic parsers like Alpino can't handle this yet, a very time consuming job awaits those who wish to compile this data, since it has to be done by hand. Texts analysed in a less sophisticated way, where words are simply tagged according to their part of speech, might help here, since it then becomes possible to more easily search for word strings that signal a certain type of coordinated ellipsis.

Another path leads down the improvement of Alpino. Hopefully it is possible to include my suggestions in the program's code and thus improve its performance in finding ellipsis. The fact that conjunction reduction is parsed correctly means that the parser is able to construe parses based on incomplete input through the use of indexing, and I think that writing specific rules that correctly handle at least certain cases of right node raising and gapping must be possible.

Finally, further research should be conducted to see if the definitions of coordinated ellipsis I postulated in chapter 2 also hold in languages other than English and Dutch. I already started to make the definitions more general in order for them to apply to both SVO- and SOV-word order languages, but maybe they need to become even more general to apply to other base-form languages. Then again, maybe it will turn out that different definitions are needed because it is impossible to unite all languages under one rule of ellipsis.

7. CONCLUSION

Although, due to the inherit difficulty of parsing ellipsis automatically in Dutch, it proved to be impossible to obtain large scale frequency data, this paper provides a useful discussion on the nature and definition of ellipsis, based in part, for Dutch, on manual corpus research. I have shown that older definitions fall short on certain aspects when they are applied to a language with word order rules different from English, such as Dutch. By piecing together these older definitions and my own findings on ellipsis I was able to postulate new definitions for coordinated ellipsis. These type-bound definitions are better able to clearly define Conjunction Reduction, Gapping, and Right Node Raising than the old site-bound definitions.

The results I did find (mainly through manual research) supported Meyer's findings. Both he and I found that CR is by far the most common type of ellipsis, followed at a distance by gapping and the rare RNR. Both English and Dutch writers used ellipsis more often than their speaking counterparts, even though the number of opportunities to use ellipsis in speech isn't significantly lower in speech. Though the numbers are higher for Dutch, they seem to confirm Meyer's arguments of Natural Flow of Speech and Semantically Less Dense Discourse, which means that the Rule of Speaker's Economy of Effort looks like a misguided way to explain why ellipsis is used. This paper shows that there are enough other reasons to use ellipsis that don't clash with the premises of Meyer's promising arguments. In particular, the Rule of Speaker's Economy of Output predicts that ellipsis will be used to produce as short of an output as possible, taking into consideration the rules that discourage the use of ellipsis. This results in more ellipsis in writing, since Meyer's arguments don't have as much impact there.

In conclusion we could say that, though ellipsis plays only a small role in speech, it is a very important role. Everyday conversation is rife with ellipsis, and it is vital to understand it if we are ever to build a perfect automatic interpreter of natural language.

8. LITERATURE

Bouma, Van Noord, and Malouf. 2001. *Alpino: Wide-coverage Computational Analysis of Dutch;* in W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavrel (eds.), *Computational Linguistics in the Netherlands 2000.* Rodolpi, Amsterdam/New York. Pages 45-59.

Carlson, G. 1977. *Reference to Kinds in English.* PhD Dissertation, University of Massachusetts, Amherst.

Dougherty, R. 1970. *A Grammar of Coordinate Conjoined Structures: I.* Language 46. Pages 850-898.

Féry, Caroline and Hartmann, Katharina. 2005. *The Focus and Prosodic Structure of German Right Node Raising and Gapping*; in The Linguistic Review, Volume 22, Issue 1. Mouton de Gruyter, Berlin/New York. Pages 69-116.

Hankamer, Jorge, and Sag, Ivan. 1976. *Deep and Surface Anaphora*; in S. J. Keyser (ed.), *Linguistic Enquiry, Volume 7, Number 3.* MIT Press, Cambridge. Pages 391-426.

Hansen, Tara. 2005. *Auditory and Visual Electrophysiological Correlates of the Processing of Gapping Structures in Adults.* Unpublished Master of Science thesis. Brigham Young University.

Harbusch, Karin & Kempen, Gerard. 2006. *ELEIPO: A module that computes coordina-tive ellipsis for language generators that don't*; in the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento.

Haspelmath, Martin. 2004. *Coordinating Constructions: An Overview;* in M. Haspelmath (ed.), *Coordinating Constructions.* Benjamins, Amsterdam. Pages 3-39.

Hendriks, Petra. 2004. *Coherence relations, ellipsis, and contrastive topics.* Journal of Semantics 21:2, pp. 133-153.

Hendriks, Petra and Spenader, Jennifer. 2005. *Why Be Silent, Some Functions of Ellipsis in Natural Language*; in J. Spenader and P. Hendriks (eds.), *Proceedings of the ESSLLI '05 Workshop on Cross-Modular Approaches to Ellipsis.* Heriot Watt University, Edinburgh. Pages 29-36.

Hoeksema, Jack. To appear. *Cumulative Effects in the Evaluation of Pseudogapping*; in J. Spenader and P. Hendriks (eds.), *Journal of Research on Language and Computation.*

Höhle, Tilman. 1983. *Subjektlücken in Koordinationen.* Unpublished manuscript, Tübingen.

Hudson, Richard A. 1976. *Conjunction Reduction, Gapping, and Right Node Raising*; in *Language, Volume 52, Number 3.* Pages 535-562.

Jackendoff, Ray S. 1971. *Gapping and Related Rules*; in *Linguistic Enquiry, Volume 2, Number 1.* Pages 21-35.

Kennedy, Chris. 2001. *Ellipsis and Syntactic Representation*; in K. Schwabe and S. Winkler (eds.), *The Syntax-Semantics Interface: Interpreting (Omitted) Structure.* John Benjamins, Amsterdam.

Koster, Jan. 1975. *Dutch as an SOV Language*; in A. Kraak (ed.), *Linguistics in the Netherlands 1972-1973*. Van Gorcum, Assen. Pages 165-177.

Levin, Nancy. 1985. *Main-Verb Ellipsis in Spoken English*. Garland Press, New York.

Lobeck, Anne. 1995. *Ellipsis: Functional Heads, Licensing, and Identification*. Oxford University Press, New York.

McShane, Marjorie J. 2005. *A Theory of Ellipsis*. University Press, Oxford.

Meyer, Charles F. 1995. *Coordination Ellipsis in Spoken and Written American English*; in *Language Sciences 17*. Pages 241-69.

Meyer, Charles F. 2002. *English Corpus Linguistics, An Introduction*. University Press, Cambridge.

Ning Yu. 2003. *Chinese Metaphors of Thinking;* in N. Salmond (ed.), *Cognitive Linguistics volume 14, issue 2/3*. Mouton de Gruyter, Berlin/New York. Pages 141-165.

Oirsouw, Robert R. van. 1984. *Accessibility of Deletion in Dutch;* in P.A.M. Seuren AO (eds.), *Journal of Semantics volume 3*. Foris Publications, Dordrecht. Pages 201-227.

Partee, Barbara and Rooth, Mats. 1983. *Generalized Conjunction and Type Ambiguity*; in R. Bäuerle, C. Schwarze, and A. von Stechow (eds.), *Meaning, Use and Interpretation of Language*. Walter de Gruyter, Berlin. Pages 361-383.

Postal, Paul. 1974. *On Raising – An Inquiry into One Rule of English Grammar and Its Theoretical Implications*. MIT Press, Cambridge, Massachusetts.

Ross, John R. 1970. *Gapping and the Order of Constituents*; in M. Bierwisch and K.E. Heidolph (eds.), *Progress in Linguistics*. Mouton, The Hague/Paris. Pages 249-259.

Sanders, Gerald. 1977. *A Functional Typology of Elliptical Coordinations*; in F. Eckman (ed.), *Current Themes in Linguistics*. John Wiley, Washington DC. Pages 241-70

Van der Beek, Bouma, Malouf, and Van Noord. 2002. *The Alpino Dependency Treebank*; in M. Theune, A. Nijholt, and H. Hondorp, *Computational Linguistics in the Netherlands 2001*. Rodopi, Amsterdam/New York. Pages 8-22.

Yatabe, Shûichi. 2001. *The Syntax and Semantics of Left-Node Raising in Japanese*; in Dan Flickinger and Andreas Kathol (eds.), *Proceedings of the 7th International HPSG Conference, UC Berkely (22-23 July, 2000)*. CSLI Publications.

Zipf, G. K. 1949. *Human Behavior and the Principle of Least- Effort*. Addison-Wesley, Cambridge, MA.

APPENDIX A - LINKS

[a] Dutch CLEF Corpus, Treebanks from the Algemeen Dagblad, year 1994 & 1995 (http://ziu.let.rug.nl/~vannoord/trees/Treebank/Machine/clef/)
[b] Online Alpino demo (http://ziu.let.rug.nl/vannoord_bin/alpino)
[c] Perl (http://www.perl.com/)
[d] Spoken Dutch Corpus/Corpus Gesproken Nederlands (CGN) (http://lands.let.kun.nl/cgn)
[e] Spoken Dutch Corpus (CGN), Treebanks from the Corpus Gesproken Nederlands (http://ziu.let.rug.nl/~vannoord/trees/Treebank/Machine/CGN/)
[f] Wordnet (http://wordnet.princeton.edu/)
[g] XPath standard (ref: www.w3.org/TR/xpath)

APPENDIX B – CORPUS SELECTIONS

From the CGN I used the following parts (found at link [e] in appendix A).

comp-b/nl/fn000004 through comp-b/nl/fn000006
comp-b/nl/fn000082 through comp-b/nl/fn000085
comp-b/nl/fn000087 through comp-b/nl/fn000088
comp-b/nl/fn000090 through comp-b/nl/fn000091
comp-b/nl/fn000092 (__2 through __5 only)
comp-g/nl/fn000069
comp-g/nl/fn000076
comp-g/nl/fn000079
comp-i/nl/fn000001
comp-i/nl/fn000010
comp-i/nl/fn000014 through comp-i/nl/fn000015
comp-i/nl/fn000018 through comp-i/nl/fn000019
comp-i/nl/fn000024 through comp-i/nl/fn000030
comp-i/nl/fn000032
comp-i/nl/fn000035
comp-i/nl/fn000046 through comp-i/nl/fn000048
comp-k/nl/fn000003
comp-k/nl/fn000033 through comp-k/nl/fn000034
comp-k/nl/fn000037 through comp-k/nl/fn000038
comp-k/nl/fn000040
comp-k/nl/fn000051 through comp-k/nl/fn000054
comp-l/nl/fn000009
comp-l/nl/fn000011 through comp-l/nl/fn000013
comp-l/nl/fn000016
comp-l/nl/fn000020
comp-l/nl/fn000036
comp-l/nl/fn000041
comp-l/nl/fn000049
comp-n/nl/fn000057
comp-n/nl/fn000060
comp-n/nl/fn000064 through comp-n/nl/fn000068
comp-n/nl/fn000070 through comp-n/nl/fn000075
comp-n/nl/fn000077 through comp-n/nl/fn000078
comp-n/nl/fn000080

From the Clef corpus I used the following parts (found at link [a] in appendix A).

AD19940103 through AD19940107

APPENDIX C – SAMPLE XML FILE

A sample XML file of the sentence "*Jan drinkt melk*" ("*Jan drinks milk.*").


```
<alpino_ds version="1.1">
  <node begin="0" cat="top" end="4" id="0" rel="top">
    <node begin="0" cat="smain" end="3" id="1" rel="--">
      <node begin="0" end="1" frame="proper_name(both,'PER')" id="2"
        neclass="PER" num="both" pos="name" rel="su" root="Jan" word="Jan"/>
      <node begin="1" end="2" frame="verb(hebben,sg3,transitive)" id="3" infl="sg3"
        pos="verb" rel="hd" root="drink" sc="transitive" word="drinkt"/>
      <node begin="2" end="3" frame="noun(de,mass,sg)" gen="de" id="4" num="sg"
        pos="noun" rel="obj1" root="melk" word="melk"/>
    </node>
    <node begin="3" end="4" frame="punct(punt)" id="5" pos="punct" rel="--"
      root="." special="punt" word="."/>
  </node>
  <sentence>Jan drinkt melk .</sentence>
  <comments>
    <comment>Q#1|Jan drinkt melk .|1|1|-0.024106819999999998</comment>
  </comments>
</alpino_ds>
```

APPENDIX D – tally_verbs.pl

This perl script lets the user count the number of times each verb in an Alpino-parsed corpus is used intransitively and the number of times it is used transitively. The script needs an input file named vl_$corpus, where $corpus is the name of the corpus to be used, which displays the location of each XML-tree that needs to be checked. My research used lists compiled with an XPath search for all sentences which contained a verb. The user has the choice to compile a list of all verbs or only those verbs used predominantly in a transitive way.

```
#!/usr/bin/perl -w
# Martijn Hennink (1158171)
#
# This script tallies the times a verb occurs transitive and intranstitive

# BODY

print "\nWelk corpus moet getallied worden? (clef/cgn)\n";

$corpus = <STDIN>;
chomp($corpus);

print "\nAlle verbs of alleen de voornamelijk transitieve? (all/trans)\n";

$mode = <STDIN>;
chomp($mode);

print "\nPlease wait while the verbs are being tallied!\n";

$loc = "/users2/s1158171/eindproject/vl_" . $corpus;

$out = "vl_" . $corpus . "_" . $mode . ".txt";

open(OUTPUT, ">$out");

if ($out eq "all") {
  print OUTPUT "\nVerbs occurring both transitive and intransitive in the " . $corpus. "
corpus.\n";
}
else {
  print OUTPUT "\nVerbs occurring mostly transitive in the ". $corpus. " corpus.\n";
}

close(OUTPUT);

open(OUTPUT, ">>$out");

open(VLIST, "$loc") or die "Error! Verblist not found!\n";
while (<VLIST>) {
  if ($_ =~ /^(\/.+xml).+$/) {
    open (TREE,"$1") or die "Error: Bestand onvindbaar!\n";
  }
  tally();
  close(TREE);
}

close(VLIST);
```

```
print "\nTally complete, data will be stored in vl_" . $corpus . "_" . $mode. ".txt!\n";

foreach $verb (sort (keys (%verbs))) {
  $nverbs = keys %{$verbs{$verb}};
  if ($nverbs > 1) {
    if ($mode eq "all") {
      print OUTPUT "\nVerb: \"$verb\"\n";
      foreach $tag (sort (keys %{$verbs{$verb}})) {
        printf OUTPUT "%-50s: %6ix\n", $tag, $verbs{$verb}{$tag};
      }
    }
    else {
      if (($verbs{$verb}{"transitive"}/$verbs{$verb}{"intransitive"} > 3) &&
($verbs{$verb}{"intransitive"} > 1)) {
        print OUTPUT "\nVerb: \"$verb\"\n";
        foreach $tag (sort (keys %{$verbs{$verb}})) {
          printf OUTPUT "%-50s: %6ix\n", $tag, $verbs{$verb}{$tag};
        }
      }
    }
  }
}

close(OUTPUT);

print "\nOutput ready! Have a nice day ;\)\n";

# SUBROUTINES

sub tally {
  while (<TREE>) {
    if ($_ =~/^.+pos="verb".+root="([a-zA-Z]+)".+sc="((in)?transitive)".+$/) {
      ++$verbs{$1}{$2};
    }
  }
}
```