

The COREA-project
Manual for the annotation of coreference in Dutch
texts

Gosse Bouma* , Walter Daelemans[†] , Iris Hendrickx[†] ,
Véronique Hoste[†] , Anne-Marie Mineur*

August 16, 2007

*Informatiekunde, Rijksuniversiteit Groningen

[†]Departement Taalkunde, Universiteit Antwerpen

Contents

1	Introduction	2
2	What is coreference?	3
2.1	Introduction	3
2.2	Definitions	4
2.2.1	Anaphor	4
2.2.2	Coreference	5
2.2.3	Antecedent / referent	5
2.2.4	Anchor point	5
2.3	Types of coreference	6
2.3.1	Identity or strict coreference	6
2.3.2	Part/whole coreference	6
2.3.3	Type-token coreference	7
2.3.4	Time-indexed coreference	7
2.3.5	Metonymy	8
2.3.6	Possessive relations	8
2.3.7	Bound anaphora	9
2.3.8	Predicate nominals	9
2.3.9	Appositions	10
2.3.10	Modality and negation	10
2.4	Syntactic categories	11
2.4.1	Pronouns	11
2.4.2	Noun phrases	16
2.4.3	Phrases without a head noun	18
2.4.4	Discontinuous NPs	19
3	About corpus annotation	20
3.1	Introduction	20
3.2	Related Annotation Work	21

3.3	The COREA project	22
4	The annotation of coreference in Dutch texts	24
4.1	Introduction	24
4.2	Types of coreference	24
4.2.1	Identity or strict coreference	24
4.2.2	Part/whole coreference	25
4.2.3	Type-token coreference	26
4.2.4	Time-indexed coreference	27
4.2.5	Metonymy	27
4.2.6	Possessive relations	28
4.2.7	Bound anaphora	29
4.2.8	Predicate nominals	30
4.2.9	Appositions	30
4.2.10	Modality and negation	31
A	Preprocessing text corpora	33
A.1	DCOI	33
A.1.1	Issues	36
B	Software	37
B.1	MMAX	37
B.1.1	Configuration	37
B.1.2	Annotation	37

Chapter 1

Introduction

This manual describes how to annotate coreference in Dutch texts. It explains what coreference is, what types there are, what terminology we use, and how we encode coreference. The manual is therefore divided into four chapters: a general chapter about coreference, a chapter about annotating coreference, a chapter that describes our implementation, and finally a chapter with examples of coreference annotation in XML.

Chapter 2

What is coreference?

2.1 Introduction

Coreference in its simplest form is about two expressions referring to the same object in the world.

- (2.1) [Xavier Malisse]₁ heeft zich geplaatst voor de halve finale in Wimbledon. [De Vlaamse tennisser]₁ zal dan tennissen tegen een onbekende tegenstander.
(*English: Xavier Malisse has qualified for the semi-finals at Wimbledon. The Flemish tennis player will play against an unknown opponent at that occasion.*)

In this example (due to [Hoste, 2005, p.200]), two noun phrases are used to refer to the same person: *Xavier Malisse* and *de Vlaamse tennisser*. The coreference between the anaphor *de Vlaamse tennisser* and the antecedent *Xavier Malisse* is based on the compatibility of the syntax and the compatibility of the content and context (world knowledge). A similar relation to the antecedent can be established using pronouns. In this example, a pronoun would have been equally suitable, cf. example 2.2.

- (2.2) [Xavier Malisse]₁ heeft zich geplaatst voor de halve finale in Wimbledon. [Hij]₁ zal dan tennissen tegen een onbekende tegenstander.

Annotating this type of coreference is relatively easy: we mark constituents with square brackets, we tag coreferential constituents with the same subscript, and the text is ready for further processing. We can collect all the information about each of the distinct referents in the text, for instance.

However, since things are rarely simple, we must also account for cases which are not strictly coreferential. This chapter discusses those.

In section 2.2, we present an overview of the definitions that are used in this manual. After that, in section 2.3, we discuss the various kinds of coreference that exist, and illustrate them with examples. In section 2.4 we discuss the kinds of syntactic categories that can act as *markables* — those categories that are relevant for annotation.

2.2 Definitions

2.2.1 Anaphor

Anaphora comes from the Greek ‘ana pherein’, *to bring back*. In other words: it brings back to mind something that was mentioned before. We use anaphora to describe lexical elements that bring back in mind an earlier element in the text.* Anaphora can be nominal as well as verbal. In this manual, we will only look into nominal anaphora. The first main division in nominal anaphora is a division into *pronouns* and *full noun phrases*.

Pronouns are abbreviated references to some entity or entities in the discourse. They contain little information, only as much as is needed to identify the intended antecedent. Pronouns can be considered pointers to antecedents.

Full noun phrases on the other hand have descriptive content that works as a criteria to single out a certain subset from a larger whole. Whereas pronouns can only impose restrictions on number and gender, full noun phrases express properties. Speaker choose those properties that uniquely describe the intended antecedent.

Given that difference, it is more difficult for a pronoun to be anything other than coreferential than it is for a full noun phrase. We shall see more on this in section 2.4.

Names are a different class of nominal phrases. Rather than refer to an element in the discourse, they refer to a deictic element. Even with repeated use, they have a constant referent in the immediate situation. In practice, they can be interchanged with pronouns or full noun phrases. Although for-

*Note that Van Deemter and Kibble [van Deemter and Kibble, 2000] use *anaphora* in a different sense. Rather than using it as a term that describes a lexical element, they use it as a term that describes a relation between two lexical elements. In their view such a relation can be either *anaphoric* or *coreferential*. Apart from *coreference* we distinguish several other relations (*metonymy*, *subset*, *bridging*). We could use the term *anaphoric* to group those, but we will avoid it, in order to avoid confusion.

mally they are not anaphora, in practice that does not make much difference in annotating, since they do refer to the same referent.

2.2.2 Coreference

Coreference is about two expressions referring to the same entity, either within the discourse (anaphora) or within the immediate situation (deixis). As we saw in the previous section, there are roughly two types of anaphora to do that: pronouns and full noun phrases. Since pronouns have little means to do otherwise, they will typically refer exactly to their antecedent, and inherit all their contents from that antecedent. Exceptions are conceivable, of course, take example 2.3.

(2.3) [The doctor]₁ entered the examination room.
 [She]₁ shook my hand.

Here, the pronoun adds information about the antecedent that we did not have previously: the doctor is female. Still, this construction only adds information, it does not contradict it, nor does it change the focus of the discourse. If that is what the speaker wishes to do, he can do so much more easily with a full noun phrase. This can go very far, as is shown in example 2.4.

(2.4) I called [the doctor]₁ to check my broken ankle.
 [The house call]₂ cost me less than I expected.

Here, *the house call* relates to *the doctor*, but it is not identical to it.

2.2.3 Antecedent / referent

That which an anaphor corefers with is called the *antecedent*. The anaphor can typically carry semantic content because of its own intrinsic meaning, or it can carry over semantic content from the antecedent it refers to. Having thus acquired semantic content, an anaphoric expression can function as an antecedent, too.

2.2.4 Anchor point

Strict coreference implies identity between an anaphor and its antecedent. However, anaphora do not always repeat an entire referent. They may refer to part of mentioned referent, or to a related element, and in that case, we cannot say that it is literally an *antecedent* anymore. However, referents,

whether identical or not, can still act as an *anchor point* in the context, determining scope. For that reason we shall generalize over coreferential and non-referential referents as *anchor points*. See also example 2.4 in §2.2.2.

2.3 Types of coreference

We distinguish the following types of reference.

2.3.1 Identity or strict coreference

We already mentioned identity, i.e. two references to the exact same object, cf. examples 2.1 and 2.2, repeated here for convenience.

- 2.1 [Xavier Malisse]₁ heeft zich geplaatst voor de halve finale in Wimbledon.
[De Vlaamse tennisser]₁ zal dan tennissen tegen een onbekende tegenstander.
- 2.2 [Xavier Malisse]₁ heeft zich geplaatst voor de halve finale in Wimbledon. [Hij]₁ zal dan tennissen tegen een onbekende tegenstander.

As pointed out in §2.2.2, we typically use pronouns to indicate straightforward identity with a previous referent. In order to add more semantic content a full, descriptive noun phrase can also be used. This is necessary when two equally suitable antecedents are available, but it can also be done simply to offer more information to the reader. Example 2.1 is a typical example of newspaper style.

Note that expressions like *eerstgenoemde*, *laatstgenoemde* and *het voorgaande* (the former, the latter, the previous) may look like full noun phrases, but in fact they behave exactly like pronouns. Their semantic content operates purely on the textual level. We could consider them verbalized pronouns.

2.3.2 Part/whole coreference

Another example of a partial coreference relation is when reference is made to a subpart of an object that was already mentioned in the discourse. These can typically take two forms: reference to members of a set, or reference to components of an object.

- (2.5) In de Raadsvergadering is het vertrouwen opgezegd in [het college]₁. In een motie is gevraagd aan [alle wethouders]₂ hun ontslag in te dienen. *English: In the council meeting the confidence in [mayor-and-aldermen]₁ has been withdrawn. A motion requests that [all aldermen]₂ resign.*
- (2.6) Hij kon [zijn auto]₁ niet meer starten. [De benzinetank]₂ was leeg. (*English: He could not get his car to start. The gas tank was empty.*)

Example 2.5 is an instance of reference to members of a set, 2.6 demonstrates reference to a component of an anchor point.

2.3.3 Type-token coreference

Type-token coreference occurs in so-called *paycheck pronouns*, a term which is named after an example from [?].

- (2.7) The man who gave [his paycheck]₁ to his wife was wiser than the man who gave [it]₁ to his mistress.

In this case, *his paycheck* and *it* do not refer to the same object in the world. So there is no identity relation between both NPs, but a semantic overlap. *It* refers to a different object, but with the same meaning.

2.3.4 Time-indexed coreference

Two expressions may refer to the same object in the world only at a particular point in time. Take example 2.8, for instance, due to [Hoste, 2005, p.217]).

- (2.8) [Bert Degraeve]₁, tot voor kort [gedelegeerd bestuurder]₂, gaat aan de slag als [chief financial and administration officer]₃. (*English: Bert Degraeve, until recently delegated manager, will start as chief financial and administration officer.*)

The expression *tot voor kort* expresses the temporateness of a predicate. The truthvalue of this utterance is related to the time at which the utterance was made. The predicate *gedelegeerd bestuurder* is true for *Bert Degraeve* at another point in time than the predicate *chief financial and administration officer*.

2.3.5 Metonymy

A very common way of referring to an object in the real world is by figure of speech: metonymy. Metonymy is defined as a rhetorical substitution of one thing for another based on their association or proximity. “The crown” is a common way of referring to a monarch, for instance, and in Belgium the castle of king Baudoin, Laken, is a common description of the Belgian monarchy. (This example is also due to [Hoste, 2005, p.218].)

- (2.9) [Boudewijn]₁ moest in die dagen niet lang zoeken naar kanalen om zijn macht in daden om te zetten. Het lijkt geen twijfel dat [Laken]₁ gedurende de hele periode 1960-1961 [zijn]₁ rol in de coulissen heeft gespeeld. [Het paleis]₁ is nooit veel meer, maar zeker nooit minder geweest dan [de exponent van de Belgische heersende klasse in haar conservatisme, in haar katholicisme, en met haar financieel-economische macht]₂.

English: In those days, Baudoin did not need to look long for channels to turn his power into deeds. There can be no doubt that Laken has played its role behind the screens during the entire period 1960-1961. The palace was never much more, but certainly not less than the exponent of the Belgian ruling class in its conservatism, its catholicism and its financial economical power.

Annotating this phenomenon is tricky. While the denotation of the term is basically the figurative and not the literal reading, it is hard to pinpoint what that figurative reading is. As can be seen in example 4.1, the possessive pronoun *zijn* in *zijn rol in de coulissen* agrees syntactically with its antecedent. The same happens in the English translation: we choose *its* to refer to the term *Laken*.

However, in a next sentence, choosing the proper pronoun is not trivial. Speakers will typically avoid this situation by using a full noun phrase, and either refer to the metaphoric sense or to the literal sense. Example 4.1 avoids this choice: *het paleis* maintains the geographical term, but still states properties that apply to humans, not buildings.

2.3.6 Possessive relations

Possessive relations express a relation between a possessor and a possessee.

- (2.10) [Rita]₁ sprak [[haar]₁ tegenstander]₂ ernstig toe.

English: Rita addressed her opponent sternly.

In example 2.10 the possessive *haar* refers to *Rita*, while the full noun phrase *haar tegenstander* refers to another person.

2.3.7 Bound anaphora

Rather than making statements about singular objects in the real world, we can also express properties of general categories. We refer to this type of referents as bound anaphora. An example is shown in 2.11.

- (2.11) [Iedereen die iets nieuws wil bereiken]₁ moet [zijn]₁ nek uitsteken.
English: Anyone who wants to achieve anything new, has to stick out his neck.

Example 2.12 is a comparable, normal case.

- (2.12) [Herman]₁ moet [zijn]₁ nek uitsteken.
English: Herman has to stick out his neck.

zijn in example 2.12 refers to a clear, singular referent in the real world. *zijn* in example 2.11 does not. We cannot paraphrase example 2.11 as *Anyone who wants to achieve anything new, has to stick out [anyone's neck]*. Such bound anaphora have a more complex meaning than that, and it is hard to express this in terms of coreference. See also §4.2.7.

2.3.8 Predicate nominals

Predicative noun phrases do *not* express a coreference relation, even though they are close. Two examples of a predicate relation (the first one due to [Hoste, 2005, p.215]) can be seen in example 2.13 and 2.14.

- (2.13) [Het mediabedrijf Vivendi Universal]₁ is [een sterke stijger binnen de DJ Stoxx50]₂.
(English: Media concern Vivendi Universal is a strong climber within DJ Stoxx50.)

- (2.14) [Ed Nijpels]₁ was als [voorzitter]₂ verbonden aan het Wereld Natuurfonds Nederland.

The annotation of predicates is still an open issue. MUC [Fisher et al., 1995], and ACE [NIST, 2003] annotation schemes do annotate predicates as part of a coreference relation, while [van Deemter and Kibble, 2000] and [Davies et al., 1998] argue against the annotation of predicates in coreference relations, because they claim that predicates not actually refer.

While we agree with Van Deemter and Kibble that predicatives are not coreferential, they are sufficiently close to be useful for our purposes. Predicative relations contain information about the antecedent which can be valuable for practical applications such as information extraction or summarization. We choose to follow the MUC annotation scheme and include predicative relations in our coreference annotation, but with a separate flag. See chapter 4.

2.3.9 Appositions

A special case of coreference is that of coreferential appositions. Appositions come in two flavors: repetitive and restrictive.

(2.15) Hu Jintao, de president van China, hield een toespraak voor de VN.
(*English: Hu Jintao, the president of China, held a speech to the UN.*)

(2.16) De Nederlandse bankgroep ABN-AMRO heeft over het tweede kwartaal van 2002 een nettowinst behaald van 534 miljoen euro.
(*English: In the second quarter of 2002 the Dutch bankers group ABN AMRO have produced a nett profit of 534 billion euro.*)

In example 2.15 two noun phrases are used to describe the same individual. Either noun phrase can be omitted without changing the meaning of the sentence. That is not the case in example 2.16 (due to [Hoste, 2005, p.214]). Such constituents belong together, they have a single ID. If we want to omit part of it, it would have to be the first part, *de Nederlandse bankgroep*, which is more descriptive but less uniquely referring than the second half, *ABN-AMRO*.

2.3.10 Modality and negation

If and when we decide to annotate predicate nominals, we should also annotate modality and negation, then these are typically phenomena that occur in predicates. It is obvious that A is not coreferential with B if a speaker uses negation and says that *A is not B*. But what about ‘A is not exactly the prototype of B’? There is negation, but more than negation, this construction expresses modality: A is B *to some extent*.

(2.17) Een partij als het CDA is, volgens Bert de Vries en andere prominente partijleden, tegenwoordig *nou niet direct het toonbeeld van sociale betrokkenheid*.

(*English: A political party like CDA, according to Bert de Vries and other prominent members, currently is hardly a model of social involvement.*)

Similarly, the use of words like *vaak*, or *irrealis* tense, expresses modality.

(2.18) De criminelen, *vaak genaturaliseerde Belgen*, ...

(*English: The criminals, often naturalized Belgians, ...*)

(2.19) Filip Dewinter *had wel eens* de nieuwe burgemeester van Antwerpen *kunnen worden*.

(*English: Filip Dewinter could have become the new major of Antwerp*)

In all of these examples (the latter two due to [Hoste, 2005, p.214, 216]) there is a partial or potential coreference relation.

2.4 Syntactic categories

Having discussed in some detail the different kinds of coreference, let us now focus on the markables: those syntactic categories that can either be anaphora or antecedents, or both, as is often the case. We have mentioned pronouns and full noun phrases (cf. section 2.1), but there is more to be said about this topic.

2.4.1 Pronouns

Pronouns come in several different flavors. We distinguish personal pronouns, R-pronouns, possessive pronouns, indefinite pronouns, reflexive pronouns and nominal ellipsis.

Personal pronouns

Personal pronouns are words such as *I*, *it*, *we*, but also *me*, *him*, *us* and *my*, *its*, *our*. Personal pronouns refer to things and animate creatures. In languages such as English, the neuter pronoun *it* refers to things, and the common pronouns *he* and *she* refer to males and females, respectively. In languages such as German, that distinction has little relation to the semantic content of the words. While *Frau* (woman) is feminine, *Mädchen* (girl) is neuter. Dutch is somewhere in between, with the Dutch that is spoken in Flanders tending more towards German, and the Dutch that is spoken in the Netherlands tending more towards English.

In example 2.20 *zij* refers to *regularisatieprocedure* — an example of the Flemish way of pronoun use.

- (2.20) [De regularisatieprocedure]₁ startte begin 2000. [Zij]₁ moest personen die al jaren illegaal in [België]₂ verblijven de kans geven via een regularisatie wettelijk in [België]₂ te wonen.

English: The regularization procedure was launched in the beginning of 2000. It was intended to enable persons which reside illegally in Belgium to legally live in Belgium through regularization.

Demonstrative pronouns

Demonstrative pronouns are very similar to personal pronouns. They can occur wherever personal pronouns occur and they, too, have a common and a neuter variant. The difference is that demonstrative pronouns are for pointing, either close or at a larger distance.

	common (hij/zij)	neuter (het)
here	deze	dit
there	die	dat

Example 2.21 and 2.22 show that nearness or remoteness is not always literally true. Example 2.23 shows that demonstratives can also refer to phrases or even sentences. Reference to non-nominal antecedents is beyond the scope of our project, however.

- (2.21) Belangstellenden kunnen ook zelf [een projectvoorstel]₁ indienen op voorwaarde dat [dit]₁ inpasbaar is in “The Language Blueprint”.

Those-who-are-interested can also submit a project proposal, on the condition that it fits within “The Language Blueprint”.

- (2.22) Aangeraden wordt [taxi’s]₁ per telefoon te bestellen, aangezien [deze]₁ doorgaans meer betrouwbaar zijn.

It is recommended to order taxis by phone, as these are usually more reliable.

- (2.23) [Heeft u ook een nieuwsbericht of een aankondiging]₁, dan vernemen wij [dat]₁ graag!

If you also have a news report or announcement, we should like to hear so.

Note that demonstratives can also occur as determiners, as is shown in examples 2.24–2.26.

- (2.24) Wijzigingen voor [deze databank]₁ kunt u melden via euromap@ntu.nl.
Changes for this databank can be reported via euromap@ntu.nl.
- (2.25) Bij [dat ontwerpproces]₁ werken technisch opgeleide studenten samen met ontwerpers en mediakunstenaars.
In that design proces technically educated student work with designers and media artists.
- (2.26) De toestand van [die wegen]₁ is veelal zeer slecht.
The condition of those roads is often very bad.

R-pronouns

Related to demonstrative pronouns are R-pronouns. They all have in common that they contain an R: *er* and *daar*.[†] *Er* has other functions as well, cf. example 2.30, where *quantitative* ‘*er*’ is used. The following examples are from the D-COI-corpus[‡].

- (2.27) De wet in [Duitsland]₁ voorziet in persvrijheid, [er]₁ is geen officiële censuur.
English: The law in Germany provides a free press, there is no official censorship.
- (2.28) Het reizen naar [het gebied]₁ wordt ontraden in verband met de aardbeving die [daar]₁ op 28 maart 2005 heeft plaatsgevonden.
English: Travel to the area is discouraged because of the earthquake that took place there on March 28 2005.
- (2.29) Aangeraden wordt in geheel Indonesië [publieke gelegenheden]₁ [waar]₁ veel buitenlanders komen te mijden.
Tourists are recommended to avoid in all of Indonesia public places that attract many foreigners.
- (2.30) [Salicylzuur]₁ werd al gebruikt, zelfs Hippocrates kende [er]₁ de werking van.
Salicylic acid was already being used, even Hippocrates knew [of-it] the effect.

[†] *Waar* is an R-pronoun as well, but it is either a relative pronoun (*de plaats waar ik woon*) or an interrogative pronoun (*waar woon jij?*).

[‡] See <http://lands.let.ru.nl/projects/d-coi/>

Possessive pronouns

Bare possessive constructions (*mine, yours*) in Dutch are usually not realized in the same way they are in English. In spoken language it does occur (*Mijne is groter dan jouwe!* — mine is bigger than yours!), but the standard form is to use a substantialized determiner: *de mijne, het jouwe*. These should be considered noun phrases rather than pronouns.

(2.31) Ieder het zijne.

English: To each his own.

(2.32) Ik denk er het mijne van.

English: I've got my own thoughts on this.

Possessive pronouns also occur as determiners in noun phrases. Examples of this are given in sentences 2.33–2.34, the latter due to [Hoste, 2005, p.206].

(2.33) [De Bank of Japan]₁ heeft beslist [haar]₁ rentepolitiek te behouden.

English: The Bank of Japan has decided to keep its interest rate policy.

(2.34) Maar al snel bleek dat ook [circuits van mensenhandelaren]₁ de procedure uitgekozen hadden om [hun]₁ ‘klanten’ België binnen te loodsen.

English: But soon, it became clear that also circuits of human traffickers had picked the procedure to sneak their ‘customers’ into Belgium.

All three kinds of possessive pronouns have in common that they are anaphoric as well as predicational. They express a possessive relation between an individual and some other entity.

Indefinite pronouns

An indefinite pronoun is a pronoun referring to an identifiable but not specified person or thing. An indefinite pronoun conveys the idea of all, any, none, or some. Typical indefinite pronouns are *men, je* and *ze*, but also *iedereen, iemand, niemand, sommigen* and *iets* are indefinite pronouns. An indefinite pronoun can be anaphoric, in the sense that the same unspecified person or thing can be referred to again, cf. example 2.37.

- (2.35) Opvallend is dat [men]₁ in de sociale huursector vooral kiest voor een ketel met een verbeterd rendement (VR-ketel), terwijl eigenaar-bewoners meestal de voorkeur gaven aan een HR-ketel.
English: Remarkable is that people in the social housing area prefer a improved yield central heating boiler (tr?), while owner-inhabitants usually prefer a high yield central heating boiler (tr?).
- (2.36) [Jouw kast]₁ was nog heel mooi. Waarom hebben [ze]₂ [die]₁ weggegooid?
English: Your closet was still fine. Why did they throw it away?
- (2.37) Die dingen kun [je]₁ niet zien. [Je]₁ kunt ze niet ruiken. [Je]₁ kunt ze niet aanraken. Maar ze zijn er wel.
English: These things one cannot see. One cannot smell. One cannot touch. But they are there.

Reflexive pronouns

Reflexive pronouns are pronouns that refer to the other argument of the same predicate, e.g. *ik was mezelf* (I was myself) or *hij vergist zich* (he is mistaken). Since third person is most common in newspaper text, the third person reflexive pronoun *zich* or *zichzelf* will be the most frequent form that we encounter. The examples are due to [Hoste, 2005, p.208].

- (2.38) [Passagiers van gekaapte vlucht 93 van United Airlines]₁ offerden [zich]₁ op.
English: Passengers of the hijacked flight 93 of United Airlines sacrificed themselves.
- (2.39) De komende weken wijdt [Coenen]₁, net terug uit vakantie, [zich]₁ volledig aan de Donna-evenementen.
English: During the next weeks, Coenen, who just returned from vacation, will commit himself to the Donna events.

Note that these cases are *inherent reflexives*: the subject of the predicate must refer to the object. This is different from normal reflexives, where the second argument of a predicate may or may not refer to the first argument.

- (2.40) Toen [Jan]₁ [een winnaar]₂ moest aanwijzen, koos [hij]₁ [Piet]₃/[zichzelf]₁.
English: When Jan had to point out a winner, he chose Piet/himself.

Nominal ellipsis

[?] formulates the principle of Speaker's Economy, which states that speakers try to minimize the amount of words needed. Pronouns exploit that principle, and in the case of nominal ellipsis they exploit it to a maximum: they do not repeat a given expression, but leave it out altogether. The following example is due to [Hoste, 2005, p. 15]:

(2.41) Roll out bottom pie crust and place \emptyset in 10 inch pie pan and set \emptyset aside.

For purposes of annotation, this is a very inconvenient habit, as there is nothing left to annotate. However, since a predicate that applies to an elliptic anaphor also applies to its antecedent — information we want e.g. a Question Answering system to be able to retrieve —, it may be relevant to annotate ellipsis as well.

2.4.2 Noun phrases

Noun phrases can be singled out a class of referents, using a semantically meaningful description. Noun phrases are potentially the most complex categories, with all the prepositional phrases, adverbial phrases and coordinations they can consist of.

Simple noun phrases

In its simplest form a noun phrase consists of a — possibly empty — determiner (an article or a quantifier) and a noun, with a possibility for adjectives. In Dutch, adjectives are placed between the determiner and the noun. Prepositional phrases follow the noun.

Example 2.42 shows several kinds of simple noun phrases. *De onzekere sfeer* is a Det-Adj-N construction. *Meer stabiliteit* is Det-N, and *het land* (Det-N) is a simple noun phrase inside a prepositional phrase *in het land* (P-NP), which in turn is a modifier to *interne spanningen* (Adj-N). The entire noun phrase is bracketed as follows: [_{NP}Adj-N-[_{PPP}P-NP]].

(2.42) [De onzekere sfeer]₁ maakt langzaam plaats voor [meer stabiliteit]₂, hoewel [interne spanningen in [het land]₃]₄ nog niet geheel zijn verdwenen.

English: the uncertain atmosphere is slowly giving in to more stability, although internal tensions in the country have not yet fully disappeared.

Conjoined noun phrases

Conjoined noun phrases are put together using words like *en* (and), *or* (of) and *met* (with). The most important point about conjoined noun phrases is their complexity. As an antecedent a conjoined noun phrase can be addressed both in part and as a whole. Example 2.43 is due to [Hoste, 2005, p. 205].

- (2.43) [Marc Coenen]₁ volgt [Jan Hautekiet]₂ op als nethoofd van jongerenmuziek-zender Studio Brussel. [Coenen]₁ stond 19 jaar geleden, samen met Schoukens en [Hautekiet]₂, aan de wieg van Studio Brussel.

English: Marc Coenen succeeds Jan Hautekiet as head of Studio Brussel. 19 years ago, Coenen was one of the founders of the youth music station together with Hautekiet and Jan Schoukens.

Example 2.44 shows that the proper names pick out the separate referents, whereas *ze* refers to both *Jan en Piet*.

- (2.44) We hebben gisteren [[Jan]₁ en [Piet]₂]₃ ontmoet. [Piet]₂ vertelde dat [ze]₃ op weg waren naar een concert van Helmut Lotti. [Jan]₁ had er duidelijk geen zin in.

English: Yesterday, we met Jan and Piet. Piet told us that they were on their way to a concert of Helmut Lotti. Jan apparently didn't feel like it.

Noun phrases containing relative clauses

Apart from adjectives and prepositional phrases, which we have gathered under the term 'simple noun phrase', a noun phrase can also be modified by a relative clause. There are two kinds of relative clauses: one restrictive (beperkende bijzin), the other one elaborative (uitbreidende bepaling). In Dutch the distinction is expressed in typography. Elaborative relative clauses are placed between commas, restrictive ones are not. Examples 2.45 and 2.46 have a distinct meaning. In the former the phrase *die het mooiste zong* is predicative to *de tenor*, in the latter the two terms are coreferential.

- (2.45) De tenor die het mooiste zong won de eerste prijs.

- (2.46) De tenor, die het mooiste zong, won de eerste prijs.

Example 2.45 expresses the proposition that the first prize was won by the tenor who sung best. Example 2.46, however, expresses the proposition that the first prize was won by the tenor, and this tenor sung best. In the first sentence, there might be more tenors, in the second sentence, there is only one.

Example 2.47 is an example from the D-COI-corpus. It is a noun phrase that singles out a set of referents (*de twee kandidaten die in de eerste ronde de meeste stemmen hebben gekregen*), whereas *die al jaren bevriend is met Toledo* in example 2.48 elaborates on a referent that was already established. Example 2.48 is due to [Hoste, 2005, p. 205].

(2.47) Op 5 mei volgt een tweede stemronde tussen [de twee kandidaten die in de eerste ronde de meeste stemmen hebben gekregen].
English: On the fifth of May there is a second election round between the two candidates who received the majority of the votes in the first round.

(2.48) [President Alejandro Toledo]₁ reisde dit weekend naar Seattle voor een gesprek met [Microsoft topman Bill Gates]₂. [Gates, die al jaren bevriend is met [Toledo]₁]₂, investeerde onlangs zo'n 550.000 Dollar in Peru.
English: This weekend, president Alejandro Toledo traveled to Seattle to talk with Microsoft top executive Bill Gates. Gates, who has been close friends with Toledo for years, recently invested about 550.000 Dollar in Peru.

2.4.3 Phrases without a head noun

Phrases with nominalized adjectives, infinitives, gerunds or quantifiers as heads can also enter into coreference relations. Example 2.49, from [Hoste, 2005, p 210], shows an antecedent that consists of a gerund (*het eten van 2 stuks fruit per dag*).

(2.49) [Het eten van 2 stukken fruit per dag]₁ wordt nog te weinig gestimuleerd. [Het]₁ is nochtans heel goed voor de gezondheid.
English: Eating two pieces of fruit each day is still under-stimulated. It is however very healthy.
(In English, the Dutch nominalized infinitive is translated as a gerund.)

Infinitives and gerunds in Dutch only differ in that gerunds have a determiner, and infinitives don't. The following is an infinitive.

(2.50) [Eerlijk zeggen wat je denkt] is het moeilijkste wat er is.

English: Honestly saying what you think is the hardest thing there is.

Example 2.51 contains a nominalized adjective *het volgende*. In fact, this is an adjective *volgend* that was derived from a verb *volgen*.

(2.51) Reizigers dienen zich rekenschap te geven van [het volgende].

English: Travellers should be aware of the following.

Example 2.52 is similar, except that there is no verb *onderstaan*. *Onderstaand* is derived from the verb *staan* and the preposition *onder*.

(2.52) In Iran kan, met inachtneming van [het onderstaande], worden gereisd.

English: In Iran one can, taking into consideration what follows below, travel.

Count nouns can also be used as the head of a noun phrase. Here, *de eerste onder de gelijken* does so.

(2.53) De koningen van de verschillende vorstendommen binnen dit rijk kozen uit hun midden de keizer, die [de eerste onder de gelijken] was.

English: The kings of the various kingdoms within this empire chose from theirs midst the emperor, who was the first among equals.

2.4.4 Discontinuous NPs

A final category is that of discontinuous noun phrases. They are not in themselves coreferential, but they can be a problem in the process of annotating. The meaning of the complete noun phrase may be different to the meaning of the parts, and so it is necessary to consider both.

(2.54) Ik heb [de jongens]_{1a} gezien [die jou zouden helpen]_{1b}. [Zij]₁ zagen mij niet.

English: I saw the boys who were supposed to help you. They did not see me.

The antecedent of *zij* in example 2.54 is *de jongens die jou zouden helpen*, and not just *de jongens* (or *die jou zouden helpen*).

Chapter 3

About corpus annotation

3.1 Introduction

The 10 categories that we discussed in §2.3, as well as the category of *discontinuous noun phrases* that we discussed under the header of *syntactic categories*, are repeated below. They are a heterogeneous group of phenomena, and they are candidates for annotation.

1. Identity or strict coreference
2. Time-indexed coreference
3. Type-token coreference
4. Modality and negation
5. Part/whole coreference
6. Predicate nominals
7. Appositions
8. Bound anaphora
9. Metonymy
10. Possessive relations
11. Discontinuous NPs

MUC-6 and MUC-7 [Hirschman et al., 1997] and [Davies et al., 1998] address items 1 (identity), 6 (predicate nominals), 7 (appositions), 8 (bound anaphora), 10 (possessive relations) and 11 (discontinuous noun phrases).

Hoste [Hoste, 2005] extends this set, and implements items 2 (time-indexed coreference), 3 (type-token), 4 (modality and negation), 5 (part-whole) and 9 (metonymy).

The COREA project includes all of these categories, and modifies some of them.

Reviewing the list helps us make an inventory of what we need. First of all, we need to be able to generalize over time. Coreference is generally a temporary issue, and even within a text we sometimes need to be able to acknowledge that. Second, we need to be able to distinguish between the sense and the reference of an expression. For type-token coreference (3), but also for bound anaphora (8) and for metonymy (9).

So how is this done? In the next sections, we describe the technical bit: the formal notation, including attributes and XML-notation.

3.2 Related Annotation Work

We can find several coreference annotation schemes for the English language, for example [Fisher et al., 1995], [Hirschman et al., 1997], [NIST, 2003], [Davies et al., 1998], [Passoneau and Litman, 1997] and [Tutin et al., 2000]. For Dutch, we have the annotation guidelines developed by [Hoste, 2005] which we use as the basis of the guidelines described in chapter 4.

Here we give an overview of the annotation scheme of [Hoste, 2005]. This scheme is largely based on the MUC-6 [Fisher et al., 1995] and MUC-7 [MUC-7, 1998] annotation scheme for English. Coreference relations are annotated using SGML tagging within the text stream. That looks as follows:

```
(3.1) Een week eerder had <COREF ID="202" TYPE="IDENT"
      REF="208" MIN="Jacques Chirac">huidig president Jacques
      Chirac </COREF> <COREF ID=209 TYPE="IDENT"
      REF="208"> zijn</COREF> kandidatuur met veel meer
      vlagvertoon bekend gemaakt.
```

The annotation scheme has the following attributes:

TYPE indicates the type of relationship that exists between anaphora and antecedents. Hoste describes four types of coreference relations: IDENT, ISA, BOUND and MOD. IDENT indicates identity relations as described in section 2.3.1. ISA is an abbreviation of *identity of sense* and refers

to the paycheck pronouns mentioned in section 2.3.3. **BOUND** indicates bound relations as described in section 2.3.7 and **MOD** refers to modality as discussed in section 2.3.10.

ID is the unique identity number of a phrase, randomly assigned.

REF contains the unique ID of the antecedent of an anaphor.

MIN is the minimum string that is the head of the phrase.

TIME is used to indicate time-dependent identity. This attribute allows for a special tag in which the temporariness of identity relations can be expressed. Except in the case of tautologies, identity is always temporary but assigning time label is not always relevant. When the identity changes within the span of time that is covered by the discourse, however, it can be essential to be able to indicate this.

3.3 The COREA project

What is new in the COREA project, is that we look at the full list of coreference-related phenomena and generalize over them. As we described in section 3.1, we need to generalize over time, and we need to be able to distinguish between sense and reference. Also, we need to be able to distinguish between different kinds of coreference: *identity*, *part/whole*, *element/set*, and *predicative*.

Our annotation scheme is largely based on the work of [Hoste, 2005]. A minor change is **MIN** becoming **HEAD**, as that is the more usual term. Contentwise, it doesn't change. The new list of attributes looks as follows.

TYPE indicates the type of relationship that exists between anaphor and antecedent. MUC-7 only has **ident**, we also have **pred** for predicative, **bound** for bound anaphora, **subset** for like elements of a set and **hasa** for components of a whole.

ID is the unique ID of a string, a numeral index that is randomly assigned by the system.

REF refers to the unique ID of the antecedent of an anaphor. Consequently, it also starts from 1 and goes upwards. This attribute is optional: a non-anaphoric element does not require a **REF** attribute.

HEAD is the minimum string that contains the meaning of the full phrase. It is a string of text.

TIME is used to indicate time-dependent identity *within a text*. A numeral index is randomly assigned by the system, to indicate that a predicate applies only at a particular point in time. Typically, an item that is marked with the TIME attribute has a temporal validity.

LEVEL is used to distinguish between the sense and the reference of a word.
Values: **sense**, **reference**.

MOD is used to express relations between phrases that are in some way modal, i.e. not quite identical. A red flag, so to speak. It can have either value **yes** or **no**. When it has value **no**, the entire attribute is typically omitted.

In the next section we see how these attributes are used in practice.

Chapter 4

The annotation of coreference in Dutch texts

4.1 Introduction

In section 2.3 we gave an overview of the different kinds of coreference that we distinguish. In this chapter we look at those examples again, and show how we annotate them.

In its most complete form a <COREF> tag has 7 potential attributes:

```
<COREF ID="237" REF="189" TYPE="ident" TIME="731" LEVEL="sense"
HEAD="man" MOD="yes">
```

However, in practice these attributes will not always co-occur, as some attributes are optional. Only ID and HEAD appear all the time. The others only apply when a constituent operates as an anaphor, and TIME is only relevant when coreference is temporally indexed. (See also section 3.3.)

When annotating, we omit any attributes that do not carry relevant information, in order to keep things transparent.

4.2 Types of coreference

4.2.1 Identity or strict coreference

Example 2.1 contains a proper name (ID 1) and a noun phrase (ID 2) that refer to exactly the same discourse referent. That relation is expressed by the TYPE attribute, which has value `ident`.

- (2.1a) [Xavier Malisse]₁ heeft zich geplaatst voor de halve finale in Wimbledon.
 (*English: Xavier Malisse has qualified for the semi-finals at Wimbledon.*)

Xavier Malisse:

<COREF ID="1" HEAD="Xavier Malisse">

- (2.1b) [De Vlaamse tennisser]₂ zal tennissen tegen een onbekende tegenstander.
 (*English: The Flemish tennis player will play against an unknown opponent.*)

de Vlaamse tennisser:

<COREF ID=2" REF="1" TYPE="ident" LEVEL="reference" HEAD="tennisser">

Example 2.2 is similar to 2.1, save for the use of a pronoun in the b-sentence. The antecedent is the same (REF=1), the relation is the same (REL=ident), the differences are in the category (CAT), a change in the HEAD value (hij instead of tennisser), and the deletion of the value for DET.

- (2.2b) [Hij]₁ zal tennissen tegen een onbekende tegenstander.

hij:

<COREF ID=2" REF="1" TYPE="ident" LEVEL="reference" HEAD="hij">

4.2.2 Part/whole coreference

In example 2.5 *alle wethouders* (all aldermen) refers to a subset of *het college* (the court — of mayor and aldermen). The type of this kind of coreference is subset.

- (2.5) In de Raadsvergadering is het vertrouwen opgezegd in [het college]₁.
 In een motie is gevraagd aan [alle wethouders]₂ hun ontslag in te dienen.
English: In the council meeting the confidence in [mayor-and-aldermen]₁ has been withdrawn. A motion requests that [all aldermen]₂ resign.

het college:

<COREF ID="1" HEAD="college">

alle wethouders:

```
<COREF ID="2" REF="1" TYPE="subset" HEAD="wethouders">
```

Example 2.6 is similar to 2.5, only the constituents of a car (gas tank, wheels, wind screen) are not its elements, and hence the relation is that of *hasa* which we do not annotate.

- (2.6) Hij kon [zijn auto]₁ niet meer starten. [De benzinetank]₂ was leeg.
 (English: He could not get his car to start. The gas tank was empty.)

4.2.3 Type-token coreference

We only present the XML-code of *the man*, *his paycheck* and *it* in example 2.7, as those are the only relevant markables for type-token coreference.

- (2.7) [The man]₁ who gave [[his]₁ paycheck]₃ to his wife was wiser than
 [the man]₂ who gave [it]₃ to [his]₂ mistress.

The XML code for *the man* is rather trivial. *The man* is not anaphoric in this context, so there is no REF attribute for it.

the man:

```
<COREF ID="1" HEAD="man">
```

His in *his paycheck* is anaphoric, it refers to *the man*, but *his paycheck* does not have an antecedent.

his:

```
<COREF ID="2" REF="1" TYPE="ident" LEVEL="reference" HEAD="his">
```

his paycheck:

```
<COREF ID="3" HEAD="paycheck">
```

It refers to *his paycheck*, but only on the sense-level (LEVEL="sense"). Just like with time-indexed coreference, the extra attribute LEVEL operates as a modifier to the REF attribute.

it:

```
<COREF ID="4" REF="3" TYPE="ident" LEVEL="sense" HEAD="it">
```

4.2.4 Time-indexed coreference

Example 2.8 shows the complexity of time-indexed reference. Both noun phrases *gedelegeerd bestuurder* and *chief financial and administration officer* refer to *Bert Degraeve*, but he does not perform both functions at the same time. These relations are time-dependent and we mark both relations with a flag for TIME.

- 2.8 [Bert Degraeve]₁, tot voor kort [gedelegeerd bestuurder]₂, gaat aan de slag als [chief financial and administration officer]₃.
(*English: Bert Degraeve, until recently delegated manager, will start as chief financial and administration officer.*)

Bert Degraeve:

<COREF ID="1" HEAD="Bert Degraeve">

gedelegeerd bestuurder:

<COREF ID=2" REF="1" TIME="101" TYPE="pred" LEVEL="reference"
HEAD="bestuurder">

chief financial and administration officer:

<COREF ID=3" REF="1" TIME="102" TYPE="pred" LEVEL="reference"
HEAD="officer">

This TIME attribute is merely a flag that this coreference is only valid at a particular point in time. It does not contain any information about which point in time, it is a warning to the person using the annotated material to be cautious and not to refer to an antecedent that is only temporally coreferential to preceding antecedents. Therefore the number of the TIME attribute is just some random number*.

4.2.5 Metonymy

Metonyms refer to the contents of their antecedent, but with an additional inference step. When we use the term *Laken* we may refer to the building in which the Belgian king lives. In example 4.1, however, the intended meaning is not literal.

- (4.1) [Boudewijn]₁ moest in die dagen niet lang zoeken naar kanalen om zijn macht in daden om te zetten. Het lijkt geen twijfel dat [Laken]₁

*In our examples we use 101 and upwards.

gedurende de hele periode 1960-1961 [zijn]₁ rol in de coulissen heeft gespeeld. [Het paleis]₁ is nooit veel meer, maar zeker nooit minder geweest dan [de exponent van de Belgische heersende klasse in haar conservatisme, in haar katholicisme, en met haar financieel-economische macht]₂.

English: In those days, Baudoin did not need to look long for channels to turn his power into deeds. There can be no doubt that Laken has played its role behind the screens during the entire period 1960-1961. The palace was never much more, but certainly not less than the exponent of the Belgian ruling class in its conservatism, its catholicism and its financial economical power.

Metonymy is not something that is resolved or created during anaphoric reference. It is the expression that is metonymic, not the anaphoric relation that is established to it. For that reason, we assume that the metonymic reading of the term has a separate entry next to the literal reading. In this particular example *Laken* is the name of the king's palace in Belgium, and it is also a term describing the king and his entourage. Each has its own lexical entry. The same holds for *het paleis*: this, too, has a metonymic reading.

Laken refers to the king proper, or rather, the king and his entourage. What makes this a particularly tricky issue for annotating is the fact that the real referent is part of the world knowledge that is associated with the term. It is not inherited from the antecedent, but it should be considered as part of the lexical semantics. Whether or not this reading of 'Laken' should be considered its prime reading is not relevant, it is the secondary reading we are using: the king and his entourage.

Boudewijn:

```
<COREF ID="1" HEAD="Boudewijn">
```

Laken:

```
<COREF ID="2" REF="1" TYPE="ident" LEVEL="reference" HEAD="Laken">
```

het paleis:

```
<COREF ID="3" REF="1" TYPE="ident" LEVEL="reference" HEAD="paleis">
```

4.2.6 Possessive relations

Possessives function as a possessivity bridge between the noun phrase that they are part of, and the antecedent that they refer to. While the possessive

is coreferential to the antecedent, the full noun phrase may be coreferential with another referent. In example 2.10 *haar* is identical to *Rita*, but *haar tegenstander* may refer to the term *Mark Rutte*, if that had appeared in an earlier sentence.

- (2.10) [Rita]₁ sprak [[haar]₁ tegenstander]₂ ernstig toe.
English: Rita addressed her opponent sternly.

Rita:

<COREF ID="1" HEAD="Rita">

haar:

<COREF ID="3" REF="1" TYPE="ident" LEVEL="reference"> tegenstander:]

4.2.7 Bound anaphora

The anaphor *zijn* in example 2.11 refers to *iedereen die iets nieuws wil bereiken*.

- (2.11) [Iedereen die iets nieuws wil bereiken]₁ moet [zijn]₁ nek uitsteken.
English: Anyone who wants to achieve anything new, has to stick out his neck.

The group of referents denoted by *iedereen die iets nieuws wil bereiken* is not fixed, rather, it is a distributive function that should be read as *for each X with property Y ...*. The anaphor *zijn* refers to that variable *X*: each *X* has to stick out his neck. However, our tools are not sophisticated enough to achieve that feat. Rather, by claiming that there is an identity relation between *zijn* and *iedereen die iets nieuws wil bereiken*, we might as well suggest that *zijn* refers to the entire group of referents. In order to mark this singularity, we use the flag *bound*. It does not resolve the lack of sophistication, but it alerts the annotator that something is afoot.

iedereen die iets nieuws wil bereiken:

<COREF ID="1" HEAD="iedereen">

zijn:

<COREF ID="2" REF="1" TYPE="bound" LEVEL="sense" HEAD="zijn">

A comparison with example 2.12 shows an additional difference between a bound and an unbound anaphor: the one refers on the sense level, the other one to the referential level.

- (2.12) [Herman]₁ moet [zijn]₁ nek uitsteken.
English: Herman has to stick out his neck.

Herman:

<COREF ID="1" HEAD="Herman">

zijn:

<COREF ID="2" REF="1" TYPE="ident" LEVEL="reference" HEAD="zijn">

4.2.8 Predicate nominals

While the copula *is* in example 2.13 suggests identity, and hence coreference between *het mediabedrijf Vivendi Universal* and *de tweede sterkste stijger binnen de DJ Stoxx50*, the relation is in fact predicative. *De tweede sterkste stijger* applies to *Vivendi*.

- (2.13) [Het mediabedrijf Vivendi Universal]₁ is [de tweede sterkste stijger binnen de DJ Stoxx50]₂.
(English: Media concern Vivendi Universal is the second strongest climber within DJ Stoxx50.)

het mediabedrijf Vivendi Universal:

<COREF ID="1" HEAD="mediabedrijf">

de tweede sterkste stijger binnen de DJ Stoxx50:

<COREF ID="2" REF="1" TYPE="pred" TIME="101" LEVEL="reference" HEAD="stijger">

4.2.9 Appositions

Repetitive appositions such as example 2.15 contain an internal co-reference relation (*ident*).

- (2.15) Hu Jintao, de president van China, hield een toespraak voor de VN.
(English: Hu Jintao, the president of China, held a speech to the UN.)

Hu Jintao:

<COREF ID="1" HEAD="Hu Jintao">

de president van China:

<COREF ID="2" REF="1" TYPE="ident" LEVEL="reference" HEAD="president">

Hu Jintao, de president van China:

<COREF ID="3" HEAD="president">

Example 2.16 contains a restrictive apposition: *de Nederlandse bankgroep* is predicative to the real nucleus of the constituent, *ABN AMRO*. We consider it a single constituent.

- (2.16) De Nederlandse bankgroep ABN-AMRO heeft over het tweede kwartaal van 2002 een nettowinst behaald van 534 miljoen euro.
(*English: In the second quarter of 2002 the Dutch bankers group ABN AMRO have produced a nett profit of 534 billion euro.*)

de Nederlandse bankgroep:

<COREF ID="1" REF="2" TYPE="pred" LEVEL="reference"
HEAD="bankgroep">

4.2.10 Modality and negation

Terms that express a positive predicate sometimes use a modality or a negative expression. We use the MOD attribute for this type of relation.

- (2.17) Een partij als het CDA is, volgens Bert de Vries en andere prominente partijleden, tegenwoordig nou niet direct het toonbeeld van sociale betrokkenheid.
(*English: A political party like CDA, according to Bert de Vries and other prominent members, currently is hardly a model of social involvement.*)

een partij als het CDA:

<COREF ID="1" HEAD="partij">

nou niet direct het toonbeeld van sociale betrokkenheid:

<COREF ID="2" REF="1" TYPE="pred" MOD="yes" LEVEL="reference"
HEAD="het toonbeeld">

Similarly, modal expressions like *vaak* are also tagged with the attribute MOD.

- (2.18) De criminelen, *vaak genaturaliseerde Belgen*, ...
(*English: The criminals, often naturalized Belgians, ...*)

de criminelen:

<COREF ID="1" HEAD="criminelen">

vaak genaturaliseerde Belgen:

```
<COREF ID="2" REF="1" TYPE="ident" MOD="yes" LEVEL="reference"  
HEAD="Belgen">
```

The use of a modal verb cluster like *had wel eens kunnen worden* also functions to indicate a fuzzy relation between anaphor and antecedent. Here, too, the MOD flag warns the user to be very cautious — to the point of rejecting the relation altogether.

- (2.19) Filip Dewinter *had wel eens* de nieuwe burgemeester van Antwerpen *kunnen worden*.
(*English: Filip Dewinter could have become the new major of Antwerp*)

Filip Dewinter:

```
<COREF ID="1" HEAD="Filip Dewinter">
```

de nieuwe burgemeester van Antwerpen:

```
<COREF ID="2" REF="1" TYPE="pred" MOD="yes" LEVEL="reference"  
HEAD="burgemeester">
```

Appendix A

Preprocessing text corpora

A.1 DCOI

The syntactically annotated part of the Stevin/DCOI corpus consists of newspaper text, and texts intended for the general public obtained from government websites.* The material has been annotated with dependency relations according to the guidelines of the DCOI-project. We have used the same texts for creating a coreference corpus.

The corpus has been converted automatically into a format suitable for coreference annotation. In particular, so-called *markables* were created by extracting all nominal constituents that can potentially be part of a coreference relation.

A constituent is any constituent in the tree which is not the head of a larger phrase. Non-leaf nodes in the tree are always constituents, with the exception of MWU nodes, which can be the (multi-word) head of a larger phrase. In figure A.1, *Carolijn Brouwer uit Nederland*, *Min de Zille*, *Vilamoura*, and *de baas* are all constituents, but just the phrase *Carolijn Brouwer* is not a constituent.

The following constituents are extracted, and annotated as a *markable*:

- Constituents of category NP
- Constituents with part of speech tag *name* or *noun*,
- Constituents with part of speech tag *pron* which are not functioning as relative pronouns.

*Currently, we only have access to the Dutch part of the syntactically annotated DCOI corpus (about 100K words).

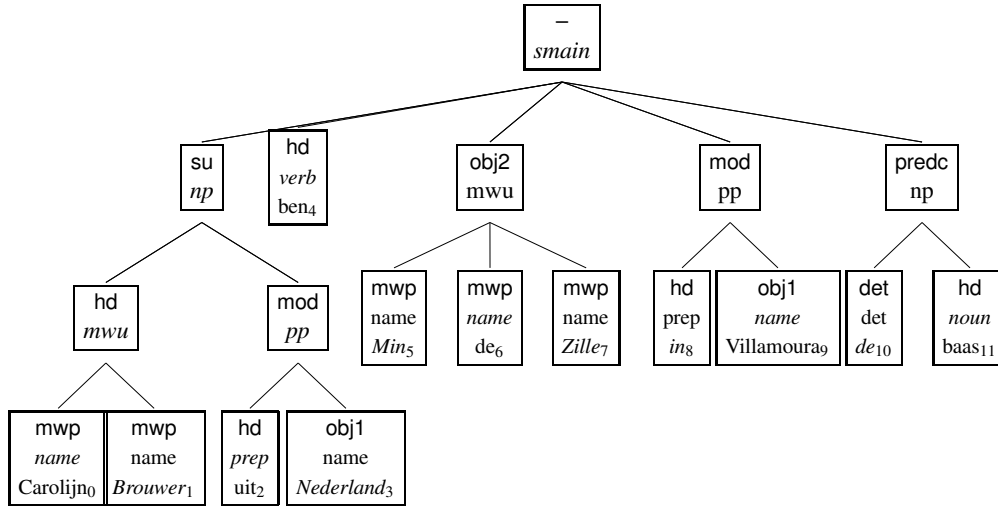


Figure A.1: Example of a DCOI dependency tree

- Constituents of category MWU which dominate a word with part of speech tag *noun* or *name*
- Constituents of category CONJ which dominate a conjunct with category NP, or part of speech tag *noun* or *name*
- Possessive pronouns, i.e. constituents with part of speech tag *det* and root form *mijn*, *jouw*, *je*, *zijn*, *haar*, *ons*, *jullie*, *uw* or *hen*
- Numerals, i.e. constituents with PoS *num*, which do not have the dependency relation *det* or *mod*.
- Determiners, i.e. constituents with PoS *det*, which do not have the dependency relation *det*.

Markables in general are in a one-to-one relation to constituents in the syntactic corpus. There are two exceptions:

- Nominal phrases containing an apposition which is a name. In these cases, the apposition is skipped as a markable. Thus, the string *zeilster Carolijn Brouwer* gives rise to a single markable for the whole NP. The embedded multi-word unit *Carolijn Brouwer* is not tagged as a markable.

En [dat alles] door [dit rotte boerenleven] ” , vertelt [Nedim Acar] . [We] bereiken [[zijn] akker] . [Hij] plukt [een handvol [graan]] , bekijkt [het] met [een zuur gezicht] en legt uit waarom [de oogst] dit jaar slecht zal zijn . ” [De hele winter] heeft [het] bijna niet geregend . [De regen] viel pas in [mei] . [Dat water] was alleen goed voor [het laten verrotten van [het graan]] . ” [De Turkse boer] is volledig afhankelijk van [de prijzen die [de overheid] betaalt voor [de landbouwproducten]] . [De slechte omstandigheden waarin [de boeren] verkeren] , zijn niet te wijten aan [die prijzen] . [De armoede op [het Turkse platteland]] wordt veroorzaakt door [[het grote aantal [boeren]] en [de technologische achterstand]] .

Figure A.2: Example of markables extracted from the DCOI-corpus.

- Nominal phrases containing a head which is a name (or a multi-word unit which contains nodes with POS=NAME) and an apposition. In these cases, the part of the constituent up to the head is tagged as a markable, and the apposition is tagged as a markable. No markable is constructed for the NP as a whole. Thus, the string *Ibrahim Yetkin , [voorzitter van de Turkse boerenbond]* gives rise to the markables *Ibrahim Yetkin* and *voorzitter van de Turkse boerenbond*.

An example of a piece of text in which the markables are shown is given in figure A.2.

Each markable in the coreference corpus has a HEAD-attribute, which is a string representing the semantic head of the constituent. The value of this attribute is obtained from the syntactically annotated corpus as well. The value of HEAD in the coreference corpus is determined as follows:

- For constituents that are phrases, whose head is a single word, it is the value of the WORD-attribute of the HD-daughter
- For constituents that are phrases, whose head is a multi-word unit, it is the concatenation of the value of the WORD-attribute of the daughters of the head
- For single-word constituents, it is the value of the WORD-attribute of the node itself

In general, the value of the HEAD attribute in the coreference corpus corresponds to the string making up the syntactic head of the constituent. There is one exception:

- For constituents containing an apposition which is a (multi-word) name, the value of HEAD in the coreference corpus is the (concatenation of the) value of the WORD-attribute in (nodes dominated by) the head node.

It should be noted that the automatic extraction and conversion process is not guaranteed to produce exactly the markables that an annotator would like to see. In those cases, markables can be added, deleted, or modified as part of the annotation-process.

A.1.1 Issues

The assignment of a *head*-value to constituents may fail in those cases where the head of a constituent is an indexed-node (which serves as a pointer to a node elsewhere in the tree). In those cases, the *head*-value will remain empty. This happens in particular in certain types of coordinations: *een Libanese en twee Syrische soldaten*. Here the head of the phrase *een Libanese* is empty.

There are a number of NP-constructions containing a modifier which is also a nominal constituent, i.e. *een aantal soldaten*, *de Ton Lutz prijs*, *drie miljard gulden*, *de Stichting Topzwemmen Amsterdam (STA)*, *Karsten Kroon (Rabo)*. Here, *soldaten*, *Ton Lutz*, *gulden*, *STA* and *Rabo* are modifiers. It seems not all of these need to be markables (*soldaten*, *gulden*, *STA*). Also, it seems in some cases it is better to consider the modifier as the semantic head. As there is no systematic way to distinguish these cases, we have not incorporated special rules for these cases. If this turns out to be problematic, the markables should be corrected manually.

Appendix B

Software

B.1 MMAX

The MMAX software that is used for annotation comes with manuals for annotators and for developers (of new annotation tasks). Below, we list the most important properties of the annotation tool, as well as a number of peculiarities which we have noticed while working with the system.

B.1.1 Configuration

- MMAX2 assumes that text is encoded in UTF-8. (That is, you can provide input in, say, ISO-8859-1 (as in the DCOI-corpus), and state this in the header of the XML files, but this information is lost after MMAX saves any updates.) If you are using a shell which does not use UTF-8 by default, you can start MMAX as

```
LC_ALL=en_US.UTF-8 java org.eml.MMAX2.core.MMAX2
```

- It is possible to define an alias (say `mmax2`), for the line above, and use this to load an `mmax` file simply by typing `mmax2 filename.mmax`
- The value of the `HEAD` attribute is stored in lower case. After MMAX saved a modified file, all your `HEAD` values will be in lower case.

B.1.2 Annotation

- Coreference relations are added by left-clicking on the markable which is coreferential. It will light up in yellow. Its `HEAD`-value will appear

in the window which displays the properties of markables. Next, select the markable to which it refers by right-clicking.

- After creation of a coreference-relation, a number of additional attributes appear in the markable-window. Choose the appropriate values for TYPE, LEVEL, TIME and MOD, and select *Apply*. (If you are happy with the defaults nothing needs to be done.)
- In general, if a selected string is member of more than one markable, a pop-up window will appear from which you can select the correct markable.
- A relation can be removed by first selecting the coreferential element with the left-button, then the element with which it is coreferential with the right button (as above), and then selecting the option *remove pointer to markable*.
- A new markable can be created by selecting a string of text with the left mouse button. After creating the markable supply a value for HEAD. Make sure no other material was already selected, otherwise you will be adding to an existing markable.
- A markable can be expanded by first selecting the markable, and then selecting additional material with the left mouse button. The new material will be added. Note that there is no requirement that the new markable must be a continuous string.
- A markable can be made smaller by selecting it, and then selecting the part to be removed with the left mouse button.
- A markable can be deleted completely by right clicking on it. Make sure no other material is already selected.

Bibliography

- [Davies et al., 1998] Davies, S., Poesio, M., Bruneseaux, F., and Romary, L. (1998). Annotating coreference in dialogues: Proposal for a scheme for mate. http://www.hcrc.ed.ac.uk/~poesio/MATE/anno_manual.htm.
- [Fisher et al., 1995] Fisher, F., Soderland, S., Mccarthy, J., Feng, F., and Lehnert, W. (1995). Description of the umass system as used for muc-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 127–140.
- [Hirschman et al., 1997] Hirschman, L., Robinson, P., Burger, J., and Vilain, M. (1997). Automating coreference: The role of annotated training data. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.
- [Hoste, 2005] Hoste, V. (2005). *Optimization Issues in Machine Learning of Coreference Resolution*. PhD thesis, Universiteit Antwerpen.
- [MUC-7, 1998] MUC-7 (1998). Muc-7 coreference task definition. version 3.0. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.
- [NIST, 2003] NIST (2003). *Entity Detection and Tracking- EDT and Metonymy Annotation Guidelines, Version 2.5.1 20030502*.
- [Passoneau and Litman, 1997] Passoneau, R. and Litman, D. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):3–139.
- [Tutin et al., 2000] Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, E., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC-2000)*, pages 28–38.

- [van Deemter and Kibble, 2000] van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in muc and related annotation schemes. *Computational Linguistics*, 26(4):629–637.