Resolving Coreferent Bridging in German Newspaper Text

von Yannick Versley

Dissertation
angenommen von der Philosophischen Fakultät
(alt: Neuphilologischen Fakultät)
der Universität Tübingen
am 19. Juli 2010

Tübingen, 2011

Gedruckt mit Genehmigung der Philosophischen Fakultät
der Universität Tübingen

Hauptberichterstatter: Prof. Dr. Erhard Hinrichs
Mitberichterstatterin: Prof. Dr. Sandra Kübler
Dekan (Zeitpunkt des Kolloquiums): Prof. Dr. Joachim Knape
Dekan (Zeitpunkt der Drucklegung): Prof. Dr. Jürgen Leonhardt

Preface and Acknowledgements

Writing a dissertation is a kind of *Bildungsroman* in itself, sketching the development of a single figure and written from this perspective – including the acceptance of a certain kind of formational ideal, interaction with the environment, moments of anguish and despair, and, finally, the cherished illusion of clarity. While written from a singular perspective, a dissertation (and the person who wrote it) owes much to, and cannot exist without, mentors, colleagues, and other kind people.

First of all, the two principal investigators of the A1 project in the SFB 441 (*"Repräsentation und Erschließung linguistischer Daten"*), Erhard Hinrichs and Sandra Kübler, deserve thanks for offering me the opportunity to pursue a PhD in Tübingen, and for playing the role of the dissertation committee. In particular, this dissertation would not have seen the light in 2010 without the benevolent and patient supervision of Erhard Hinrichs. It is due to his encouragement that this thesis is not just a cumulation of scientific facts but also offers an inviting narrative for the curious reader. Sandra Kübler, in her time in Tübingen, tirelessly listened to any problems and uncertainties that put themselves in the way, despite having lots of her own work, and always offered sensible and insightful advice.

My time in Tübingen would have been rather dull, or worse, without all the colleagues, co-workers and friends: Starting from the far end of the alphabet, Heike Zinsmeister helped me to discover that, despite all the vagaries in general and academic life, there is a purpose to it all that you should try and discover. Holger Wunsch, first my office mate and then my floor neighbour, helped me feel at home in Tübingen early on with his cheerful demeanor. Piklu Gupta, in his own way contributed not only his perspective on things, but was also very adept in listening to, and understanding other perspectives, helping create a coherent picture.

In the larger environment, lots of people have shared their perspectives on life, research, and sometimes mensa food – Kathrin Beck, Stephan Kepser, Monica Lău, Lothar Lemnitzer, Daniela Marzo, Frank Müller-Witte, Verena Rube, Ilona Steiner, Thomas Zastrow, Bettina Zeisler, among others. The members of the new SFB 833 project, namely Stefanie Simon, Sabrina Schulze and Anna Gastel, deserve thanks for enduring the anguish and brooding of a dissertation in its final stadium.

With several people outside of Tübingen – to name a few, Sabine Schulte im Walde, Stefan Evert, Marco Baroni, Massimo Poesio, Olga Uryupina, Josef van Genabith, Ines Rehbein – I had stimulating discussions, for which I am particularly grateful.

Contents

Chapter 1

Introduction

1.1 The Problem

Anaphoric descriptions (such as pronouns) and rigid designators (such as names) make it possible to connect the contents of multiple sentences in a discourse and therefore are important means for the structuring of the discourse beyond the level of single sentences. Reconstructing the sets of mentions that relate to the same entity in a discourse model by resolving anaphora and names in a text, so-called **coreference resolution**, has been proven to be useful for a wide range of higher-level tasks such as question answering (Morton, 2000), summarization (Steinberger *et al.*, 2005) and information extraction (McCarthy and Lehnert, 1995).

Consider the following stretch of discourse[1]:

(1.1) [$_1$ *President Clinton*] *had planned weeks ago to devote yesterday to building up public interest in next week's State of the Union address.*
Instead, [$_1$ *he*] *spent* [$_1$ *his*] *afternoon with a revolving door of reporters, in a campaign to keep* [$_1$ *his*] *presidency from buckling under the force of allegations about* [$_1$ *his*] *relationship with* [$_2$ *a former White House intern*].
In a remarkable series of three interviews in which [$_1$ *the president*] *was questioned bluntly and without apology about adultery and obstruction of justice alike,* [$_1$ *Clinton*] *denied having had a sexual relationship with* [$_2$ *a then-White House aide, 24-year-old Monica Lewinsky*].

"*President Clinton*", multiple occurrences of "*he*" and "*his*", "*the president*" and "*Clinton*" all designate the same person. Yet, without coreference resolution, an information extraction system would lack the information that "*the president*", who was questioned bluntly, is the same person who (allegedly) has a relationship with a White house intern. In extractive summarization, information about the discourse-old status of "*the president*" and about the previous elements of the coreference

[1] ACE-02 corpus, newspaper section, file 9801.172

chain could be used to substitute the mention with a more complete description ("*President Clinton*") if the third sentence, but not the first two, were part of the summary.

To perform coreference resolution in good quality, several linguistically interesting subproblems have to be addressed. One major group is the area of *pronouns*, while the other encompasses or full noun phrases (*non-pronouns*).

Third-person pronouns have drawn the interest of computational linguists for quite a long time (Hobbs, 1978; Lappin and Leass, 1994; Mitkov, 1998; Tetrault, 2001, *inter alia*; see Wunsch, 2010 for a modern perspective on this topic). As can be seen in the example (1.1), personal and possessive pronoun references such as *he* and *his* are quite frequent in newspaper text. Non-referential occurrences of *it* can be found in most text types and have to be filtered out since there is no coreferent antecedent.

First and second person pronouns, superficially very similar to their third-person counterparts, function differently because their content depends less on the previous utterances and more on the context of the utterance (especially, the speaker of that utterance).

First and second person pronouns frequently occur in reported speech, as in the following example (1.2):[2]

(1.2) *Declaring that " [$_1$ I] am not a crook", [$_1$ President Nixon] vigorously defended [$_1$ his] record in the Watergate case tonight and said that [$_1$ he] had never profited from [$_1$ his] public service. " [$_1$ I] have earned every cent. And in all of [$_1$ my] years of public life [$_1$ I] have never obstructed justice," [$_1$ Mr. Nixon] said.*

Demonstrative pronouns most frequently refer to events or facts, and are only very rarely coreferent with prior noun phrases in the text.[3]

(1.3) a. *There were some scuffles, there was a little bit of stone throwing and at one point [$_1$ the Western War[sic] Plaza, the area that stands in front of the Jewish holy site of the West Moor] was cleared, [$_1$ it] is adjacent to the Muslim site of the Haram al-Sharif or Temple Mount.*
b. *[$_1$ This] was cleared because of some stone throwing.*

On the side of non-pronouns, the most important distinction is that between (proper) names and nominals (noun phrases with a common noun head).

Names such as *Bill Clinton*, *Monica Lewinsky*, or *Richard Nixon* uniquely specify an entity; however, there often is substantial variation in the exact form used, such as *President Clinton*, *Clinton*, or *Mr. Clinton*. Shortened names (*Clinton*, *Mr. Clinton*) are sometimes, but relatively rarely, ambiguous.

[2] From a 1973 Washington Post article:
`http://www.washingtonpost.com/wp-srv/national/longterm/watergate/articles/111873-1.htm`

[3] Example from the OntoNotes corpus, document pri_0096.

The class of **Nominals** designates noun phrases with a common noun head (and, in annotation schemes which include them, premodifying bare nouns). Nominals that are discourse-old (i.e., subsequent mentions of an entity introduced earlier) usually come in the form of noun phrases with a definite article, such as *the president*, but also noun phrases with a demonstrative article (such as *this group of determined volunteers*). In the rest of the thesis, noun phrases with a definite article will be called *definite descriptions*. In some cases, definite descriptions are discourse-old, i.e., they corefer with a prior mention that has the same head (*a car ... the car*). In other cases, however, a discourse-old definite description has a different head (*a 1957 Ford Mustang ... the car*). In addition, slightly more than half of all definite descriptions are discourse-new: despite having a definite article, these noun phrases introduce a new entity in the discourse.[4]

Among the category of nominals, one can also find noun phrases with a possessive article, such as *his record in the Watergate case*, *his public service* or *my years of public life*, or with a demonstrative article, such as in the following example:

(1.4) a. *In June 1988* [$_1$ *ACET*] *was officially launched as a registered charity.*
b. *Over the next six months* [$_1$ *this small group of professionals and volunteers*] *provided practical care for 15 people living in or around the Borough of Ealing.*

Similarly to demonstrative pronouns, demonstrative full noun phrases are often used to refer to entities not previously introduced as a noun phrase, such as *this year*, *this process*, or *this way*.

Finally, as in example (1.1), indefinite noun phrases can be used to specify the same entity, as with "*a former White House intern*" and "*a then-White House aide, 24-year-old Monica Lewinsky*".[5]

1.2 Scope of the thesis

This thesis is concerned with techniques to improve coreference resolution for **German newspaper text** with respect to the non-pronoun part of the problem, namely the resolution of **definite descriptions** (i.e., common noun phrases with a definite article) as well as **names**.

[4] The exact number depends not only on the language – languages such as Spanish or Italian use a definite article before possessive pronouns, as in *la mia sorella* / [the] my sister – but also varies by genre. Estimates range from 53% of definite descriptions being discourse-new (Vieira and Poesio, 2000) to 68% of mentions with a definite article (Recasens and Marti, 2009). Ariel (2007) points out that the percentage of definite descriptions that is anaphoric (as opposed to uniquely describing or exophoric) shows considerable variation even across different types of written text within the same language.

[5] The fact that the text uses multiple indefinite noun phrases "*a former White House intern*" and "*a then-White House aide*" to refer to Monica Lewinsky is not a typical use and may be due to the writer intending to imply that her [exact] identity is not relevant for the adultery accusations.

Definite descriptions and names are not always discourse-old, and as they carry substantially more information than pronouns, different disambiguation strategies are necessary than for pronoun resolution.[6]

To find a previous mention for a name or definite description, it is often possible to use simple (sub-)string matching techniques, as in (1.1): sub-string matching techniques allow us to find out that "*President Clinton*" is a plausible antecedent for the descriptions "*Clinton*" and "*the president*".

Many definite descriptions, however, have antecedents that do not share any substring and which make it necessary to consider additional knowledge if accurate coreference resolution is to be performed. Among the problems in the resolution of definite descriptions, this problem of **using external knowledge** – either from existing resources or learned in an unsupervised fashion – which would allow it to resolve cases such as the following examples (1.5) and (1.6), are among the most interesting. Accordingly, the techniques that are necessary for the resolution of non-same-head definite descriptions form the centerpiece of the work presented here.

As an illustration, let us consider the two following examples:

(1.5) a. Lebensgefährliche Körperverletzungen hat sich [$_1$ eine 88jährige Fußgängerin] bei einem Zusammenstoß mit [$_2$ einem Pkw] zugezogen.
[$_1$ *An 88-year-old (female) pedestrian*] *has been gravely injured in a collision with* [$_2$ *a car*].

b. [$_1$ Die Frau] hatte [$_2$ das Auto] beim Überqueren der Waller Heerstraße offensichtlich übersehen.
When crossing the Waller Heerstraße, [$_1$ *the woman*] *had obviously overlooked* [$_2$ *the automobile*].

(1.6) a. "Wir müssen uns selbst helfen", meinte [$_3$ Magath].
"We have to help ourselves", said [$_3$ *Magath*].

b. [$_3$ Der Werder-Coach] bangt um die Einsätze von Detlef Eilts und Marco Bode.
[$_3$ *The Werder Bremen coach*] *is worried about Detlef Eilts and Marco Bode's ability to play*.

In discourse (1.5), the description "eine 88jährige Fußgängerin" (*an 88-year-old (female) pedestrian*) is taken up again with a mention of "die Frau" (*the woman*), which is strictly less specific in its logical content, whereas "ein Pkw" (*a car*) and "das Auto" (*the automobile*) are more or less equivalent terms in this context.

In (1.6), where the salient type "Werder-Coach" (*Werder Bremen coach*) is used after the name mention "Magath", we see how entities introduced with names can subsequently be mentioned with contextually salient descriptions. (Notice that "*the*

[6] In the experimental part of this thesis, full noun phrases with a demonstrative article will be included in the experimental data and treated in the same way as definite descriptions, despite the fact that the demonstrative article has a somewhat different meaning from the definite article's.

son of a Puerto Rican soldier" may also be a possible description for Felix Magath, but would be odd here as the information in that mention would neither be known to the reader nor a plausible addition in this context).

Such cases, in which shallow techniques based on string matching are not effective and where one has to approximate the semantic content of mentions somehow, are relatively frequent – only 60% of discourse-old definite noun phrases in a sample from the referentially annotated TüBa-D/Z corpus (Hinrichs *et al.*, 2005a) are resolvable to a same-head antecedent, whereas the remaining 40% can only be resolved using more elaborate knowledge sources.

The rest of this chapter deals with the relationship between coreference resolution and the related linguistic phenomena of anaphora and bridging. Chapter 2 looks at linguistic assumptions about definite descriptions which can inform the coreference resolution, whereas chapters 3 and 4 serve to give an overview about the existing literature on coreference resolution including the resolution of names and definite descriptions. Because of the applied nature of coreference resolution, relatively few articles go into the specifics of a given subproblem such as matching names or resolving definite descriptions. Instead, most researchers prefer to describe their system as a whole rather than focusing on specific components. Finally, chapters 6 and 7 present experiments on the resolution of definite descriptions and nominals: Chapter 6 evaluates existing and novel features for the resolution of definite descriptions in a simplified setting that is frequently used to corroborate the efficacy of methods for this sub-problem. Chapter 7 assumes a more holistic perspective and discusses the resolution of definite descriptions and names in a realistic setting.

1.3 Coreference and Anaphora

Let us start by clarifying the relation between the related but distinct terms of coreference and anaphora.

I will follow Halliday and Hasan (1976) as using the term **anaphor** for textual elements that need additional information to be interpreted and where that information is retrieved from the previous text (as opposed to *exophora*, which make reference to the context of the situation). The preceding text element where this information comes from is called the **antecedent**.

Halliday and Hasan's cohesion-based definition of anaphora is to be seen in contrast with narrower uses of the term in formal syntax, especially the study of anaphoric binding (Reinhart, 1983; Lasnik, 1989). Anaphoric binding has as its subjects the structural constraints on the form of noun phrases that arise from particular configurations.

Chomsky's (1981) exposition of government and binding theory establishes a different use of the word *anaphor* by proposing a classification of noun phrases into '*anaphors*' (a term which Chomsky uses exclusively for reflexives), '*pronominals*' (meaning non-reflexive pronouns), and '*R-expressions*' (non-pronominal

noun phrases). It is important to note that Chomsky's use of the term *anaphor* differs from the way that it is or was understood by Reinhart, or contemporary semanticists such as Partee (1978), but has become common use in generative syntax.

To avoid misunderstandings, it is therefore a good idea to use the unambiguous terms *anaphoric binding* and *discourse anaphora* whenever the intended meaning is not clear from the context.

In computational linguistics, the term "*anaphora resolution*" is clearly understood in Halliday and Hasan's broader sense of resolving incomplete descriptions to an antecedent – for example, the survey article of Mitkov (1999) explicitly mentions pronominal anaphora, definite noun phrase anaphora (called *definite descriptions* here), and one-anaphora (discussed as *identity-of-sense anaphora* on page 20, below). Despite this fact, there is a tendency to see third person personal pronouns as the paradigmatic example of anaphora since referring third person pronouns are nearly always anaphoric.

Definite descriptions can be anaphoric (as in *Bremen. . . the city*), but frequently occur as non-anaphoric **unique descriptions**: noun phrases such as "*the 42nd president of the United States*" or "*the first person to cross the Atlantic*" can be interpreted independently of the previous text in much the same ways as names can.

Besides denoting specific individuals, unique definite descriptions[7] can also refer to a well-established kind as in the following example (1.7) by Krifka (1987):

(1.7) A father points to lions in a zoo cage and says:
This is [$_1$ *the lion*].
[$_1$ *It*] *lives in Africa.*

The exact semantics of such generic uses is notoriously hard to pin down. As a consequence, many researchers such as Hawkins (1978) or Krahmer (1998) explicitly exclude them from their account, and an adequate treatment would be beyond the scope of this thesis.

Discourse-old definite descriptions may sometimes be anaphoric and have an anaphor-antecedent relation to the previous mention, but they can also uniquely refer to an entity already introduced in the text, in which case no anaphoric relation exists, despite the fact that the previous mention and the subsequent definite description are coreferent.

To see how a unique description may be discourse-old but non-anaphoric, consider the following example (1.8): The name "*Bill Clinton*" in (1.8-c) can be interpreted independently of the text; therefore, it is not anaphoric.

(1.8) a. [$_1$ *William Jefferson "Bill" Clinton*] *was the 42nd President of the United States from 1993 to 2001.*

[7] Generic use of incomplete definite descriptions as in "*We discussed the Siberian tiger in class. The tiger is nearly extinct.*" seems to be rather unnatural.

b. [$_1$ *He*] *became president at the end of the cold war, and as* [$_1$ *he*] *was born in the period after World War II,* [$_1$ he] *is known as the first baby boomer president.*

c. [$_1$ *Bill Clinton*] *was born William Jefferson Blythe, III, in Hope, Arkansas.*

In other cases, definite descriptions such as "*the president*" in (1.1) can be interpreted anaphorically (in relation to the previous mention "*President Clinton*"), but would also be interpretable exophorically (the current president with respect to date and location of the writing of the article).

Anaphoricity plays a large role in the coreference resolution of definite descriptions, yet it plays a different role than in pronoun resolution. The fact that coreference resolution commonly shares a large part of the algorithmic problem (finding previous mentions of the same entity) with pronoun resolution occasionally leads to confusion when it comes to describing resolution algorithms.

Some researchers (Soon *et al.*, 2001; Yang *et al.*, 2005b) use terms such as "*anaphor*" and "*antecedent candidate*" discussing the resolution of a mention that is suspected to be discourse-old, independent of the anaphoricity of that mention. Other researchers in computational linguistics carefully tread around the issue and prefer terms such as "NP_i" and "NP_j" (Ng and Cardie, 2002b, who still use '*anaphoric*' for subsequent mentions), or "*active mention*" and "*in-focus entity*" (Luo *et al.*, 2004).

To avoid confusion in this domain, yet stay reasonably close to established terminology, I will use the term **coreference antecedent** for a previous mention of the same entity; for the non-anaphoric mention "*Bill Clinton*" in example (1.8), the previous mention "*William Jefferson 'Bill' Clinton*" would be a coreference antecedent, which a system performing coreference resolution would have to find out among other (coreference) **antecedent candidates** such as "*World War II*", or "*the cold war*" which are preceding mentions of different entities.

The notion of coreference antecedent avoids the notational confusion of Soon *et al.* and Yang *et al.* (since not all noun phrases are anaphoric), but at the same time avoids the lack of intuitive clarity that comes with Ng and Cardie's NP_i and NP_j or Luo *et al.*'s *active mention*.

The approach I will present in the rest of this thesis uses a hybrid model with both rule-based and machine learning compontens, which allows to combine different knowledge sources – taxonomic information from GermaNet, semantic classes, pattern-based web searches, and distributional models – while at the same time maintaining fine-grained control over the resolution algorithms. It makes an explicit connection from postulated semantic relations to properties of the knowledge sources to motivate the particular choice of knowledge sources, and presents a way of integrating the task of selecting an antecedent with that of discovering if a definite description is discourse-old, as in examples (1.5) and (1.6), or discourse-new (in which case there is no antecedent to which it can be resolved).

1.3.1 Notational Conventions

- **Coreference resolution** designates the kind of computational or annotation task where spans of text corresponding to certain linguistic entities (most often, noun phrases) are marked up and grouped into equivalence sets according to criteria that aim at referential properties of these linguistic entities.

  This definition is deliberately broad to encompass the whole breadth of tasks including the noun phrase coreference task where noun phrases in a discourse are grouped if they are mapped to the same discourse entity in a (mental or abstract) discourse model, but also tasks such as cross-document coreference where rigid designators for each (well-known or at least named) real-world entity are grouped across texts (so-called *cross-document coreference*, see also Bagga and Baldwin, 1998b, Holler-Feldhaus, 2004).

  Coreference resolution in this sense of 'computational or annotation task' does not necessarily correspond to a single linguistic phenomenon; indeed, nominal coreference resolution is best understood as covering both anaphoric descriptions and rigid designators (see section 2.5), and some coreference schemes include prenominal modifiers as mentions (e.g. "aluminum" in "*the aluminum price*") although they are identical in sense rather than in reference, and include in the coreference definition the purely syntactic phenomena of appositions and copula constructions (see section 3.1).

- **Anaphoric descriptions** are mentions (most commonly, noun phrases) that are incomplete in that they cannot be understood without a context, and can be interpreted using the previous text, (or, in spoken language, the previous utterances).

  Third-person pronouns are a prototypical example of anaphoric descriptions in that they are almost always used anaphorically and can only be used non-anaphorically in very specific contexts (e.g. "*THEY are after you*"). However, common noun phrases can and are used anaphorically: noun phrases such as "*the men*" or "*the city*" are incomplete and usually anaphoric.

  Because this thesis is concerned with names and common noun phrases, "the anaphor" will in most contexts designate an anaphoric common noun phrase and not a pronoun.

- The class of **definite noun phrases** comprises pronouns, names, and noun phrases with a definite article or a possessive or demonstrative determiner, whereas the term **definite description** more narrowly designates noun phrases with a definite article (cf. Löbner 1985).

  Whereas Heim's account of definiteness (see section 2.2) is meant to cover all kinds of definite noun phrases, the presentation in chapter 2 is concerned with the narrower category of definite descriptions, and "*a definite*" will always mean a definite description.

- A mention is **discourse-old** whenever the entity it refers to has already been mentioned in the discourse. If it has not been mentioned previously, it is counted as **discourse-new**, even if it is previously known to the hearer (e.g., *the sun*, *Bill Clinton*).

1.4 Anaphora and Bridging

The title of this thesis makes reference to the term '*bridging*', following Vieira and Teufel (1997), whose *coreferent bridging descriptions* the title alludes to. Vieira and Teufel simply define bridging definite descriptions as "*uses of definite descriptions whose antecedents – coreferent or not – have a different head noun*". Given this relatively frugal description, it seems appropriate to elaborate the idea of bridging somewhat and to clarify its relation with the concepts of anaphora on one hand and coreference on the other hand.

Both anaphora and bridging subsume coreferent case such as in the following example (1.9-a) as well as non-coreferent cases such as (1.9-b).[8]

(1.9) a. I met a man yesterday.
The bastard stole all my money.
b. I looked into the room.
The ceiling was very high.

There is, however a difference in perspective: While Halliday and Hasan's (1976) idea of anaphoric reference focuses on indicators that signal that the speaker is to retrieve additional information from elsewhere, and the constraints on how to identify the information, Clark's (1975) notion of bridging emphasizes the role of the inferences that a listener may make, and how the inferences are constrained so that the inferencing process remains efficient and predictable.

Clark's account of bridging is that a new segment of discourse contains both **given** (or thematic) information, which is asserted to be mutually known, and **new** (or rhematic) information, which will be added to the shared mental representation that speaker and hearer build; the distinction between given and new may be marked through syntactic and/or intonational means. Clark then describes the construction of a bridge (relating given information to an antecedent) as follows:

> **Given-new contract:** The speaker agrees to try to construct the Given and New information of each utterance in context (a) so that the listener is able to compute from memory the unique Antecedent that was intended for the Given information, and (b) so that he will not already have the New information attached to the Antecedent.
>
> The listener in turn knows then, that the speaker expects him to have the knowledge and mental wherewithal to compute the intended Antecedent in that context, and so for him it becomes a matter of solving a problem.

[8] Examples 7 and 13 of Clark (1975).

> What *bridge* [emphasis added] can he construct (1) that the speaker could plausibly expected him to construct and (2) that the speaker could plausibly have intended? The first part makes the listener assess principally what facts he knows and the second what implicatures he could plausibly draw.
>
> Bridging – the construction of these implicatures – is an obligatory part of the process of comprehension. The listener takes it as a necessary part of understanding an utterance in context that he be able to identify the intended referents (in memory) for all referring expressions. (. . .)

As a consequence of the different perspectives assumed, prototypical cases of anaphora are those where the anaphor is semantically poor and the antecedent easily identifiable (as in personal pronouns), whereas prototypical cases of bridging are those where the semantic relation between anaphor and antecedent is non-trivial. The latter cases where non-trivial inferences are necessary are also called **accommodation** following Lewis (1979).

Neither Halliday and Hasan's approach nor Clark's require that the antecedent belong to the same grammatical class as the anaphoric description; In coreference annotation, however, one would normally want to avoid cases such as the following example (1.10):[9]

(1.10) Combine the flours, salt and thyme in a bowl of a food processor. Add the butter and pulse until [the mixture] resembles little pebbles in a beach of sandy flour (about 20 quick pulses).

Any competent reader of cooking recipes will notice that combining ingredients in a food processor and mixing them will result in a mixture; hence, *the mixture* is plausibly part of the common ground (i.e., given information), yet coreference resolution would treat the noun phrase as *discourse-new* since the entity has never been mentioned using a noun phrase.

Focusing on referential identity instead of assuming a more detailed view on allows for repeatable and affordable annotation of coreference in corpora, whereas the wider range of phenomena found in non-coreferent bridging anaphora often cannot be reproduced reliably.

Bridging at large includes even such anaphoric relations that are not based on the identity of the corresponding discourse entities, as in the anaphoric examples in (1.11). Moreover, some researchers such as Asher and Lascarides (1998) also see bridging implicatures at work in cases such as (1.12), where implicatures occur with no anaphoric noun phrase triggering them:[10]

(1.11) a. Peter loves Joan. [The fact] is very obvious.
b. I took my car for a test drive. [The engine] made a weird noise.

[9] `http://www.101cookbooks.com/archives/buckwheat-cheese-straws-recipe.html`

[10] (1.11-b) and (1.11-c) due to Asher and Lascarides (1998), (1.12-a) due to Hawkins (1978), (1.12-b) and (1.12-c) due to Charniak (1983)

c. I've just arrived. [The camel] is outside and needs water.
d. Peter has a red car. I want [a green one].

(1.12) a. Fred bought a book from the bookstore. [A page] was torn.
b. Jack was going to commit suicide. He got [a rope].
c. Jack got a rope. He was going to immobilize Bill.

Bridging acts as a unifying view on all these phenomena, including inferences that connect an indefinite noun phrase to the previous discourse.[11]

It is probable that examples such as (1.11-c) and the examples in (1.12), as well as less clear-cut cases that would occur in texts, would pose a real problem to both annotation and automatic resolution. This, however, should not be taken to mean that researchers have not proposed successful descriptive and theoretical accounts for bridging phenomena outside coreference, or that corpus annotation and automatic resolution for such anaphoric phenomena were a hopeless endeavour.

Abstract object anaphora (1.11-a), associative anaphora (1.11-b), and identity-of-sense anaphora such as (1.11-d) all have seen successful approaches, and researchers have successfully identified constraints on admissible anaphor-antecedent relations in these cases.

Abstract Object Anaphora The first group consists of cases such as (1.11-a), where a proposition in the discourse ("*Peter loves Joan*") is the antecedent for a noun phrase ("*The fact*"). This subset of anaphoric relations has been called *abstract object anaphora* (Asher, 1993), *complex anaphora* (Consten and Knees, 2005) or *discourse deixis* (Webber, 1988), with accounts generally agreeing that propositions (and events) can get reified (i.e., converted into entities of their own) and reference be made to them, and varying in the question of whether propositions are always reified or if this is a side-effect of a reference such as "*this*" or "*the fact*". Byron (2004), who provides a computational implementation for the resolution of abstract object anaphora in personal and demonstrative pronouns in the TRAINS domain, notes that e.g. requests (such as "Get the Oranges to Corning") enable event reference ("That will take three hours"), but not reference to a proposition ("??That's right").

Associative Anaphora The case of bridges from one noun phrase ("*the car*") to another noun phrase that is not coreferent but related ("*the engine*") is generally referred to as associative anaphora (Hawkins, 1978, p. 123). While meronymic bridges such as *car–engine* may be seen as prototypical, Asher and Lascarides (1998) express doubt whether even the extensive taxonomy of Clark (1975) (which includes necessary/optional parts as well as necessary/optional roles) is sufficient to

[11] Asher and Lascarides (1998) propose that fulfilling the given-new contract also involves computing a discourse relation, explaining the role of the rope in cases like (1.12-b) and (1.12-c). Hawkins (1978) proposes that in (1.12-a), the plausible presence of the set of pages of the book guides the inference in a similar way as it would in the case of the definite "*the title page*".

cover all such cases; they cite bridges such as "*arrive – camel*" in (1.11-c) where it is inferred that the camel has been the means of transport for the travel culminating in the arrival.

Similarly to the work on coreferent bridging that I will present in the rest of the thesis, most computational work on associative anaphora relies on lexical relations, in this case meronymic relations either from wordnets (Vieira and Teufel, 1997), from distributional association metrics (Poesio *et al.*, 1998), or by generalizing found meronymic relations with the help of taxonomic information in wordnets (Meyer and Dale, 2002).

Identity-of-Sense Anaphora In other cases, information from the discourse is used not to retrieve a specific referent, but rather a specific type, for example in *paycheck anaphora* such as (1.13), where *it* introduces a new referent but copies the type *paycheck*:

(1.13) The man who gave his paycheck to his wife is wiser than the one who gave [it] to his mistress.

Similarly, *one-anaphora* are cases where *one* is used as a place-holder for the semantic content of (part of) a previous description in the discourse.

Why 'coreferent bridging'? Throughout this thesis, I will use the term '*coreferent bridging*' to refer to cases where a definite description is a subsequent mention of an entity that has already been mentioned, but where the relation between the *information* contained in the current and prior mentions is not trivial (i.e., where none of the prior mentions has the same head noun).

Given the breadth of Clark's use of the term *bridging*, it could be argued that the term may be seen as confusing, as coreference annotation (in the narrow sense) excludes all of the cases where the relation in terms of the entities involved is non-trivial, and associative anaphora such as (1.11-b), but also more complex relations as in (1.11-c) or the examples from (1.12), may be argued to be more prototypical cases of bridging.

However, besides the obvious reason that this is a term that previous research has established (see below), it also directs our attention to the fact that the informational relations between (potentially) anaphoric definites and their (potential) antecedent is not always clear-cut.

Starting from Clark's rather broad meaning of *bridging*, including referential identity (including pronominalization) as well as associative and abstract object anaphora, Vieira and Teufel (1997) use the term *bridging references* for "*uses of DDs [definite descriptions] whose antecedents – coreferential or not – have a different head noun*", a classification which is also used in Vieira and Poesio (2000)'s work on resolving definite descriptions. Vieira and Poesio distinguish between *direct anaphora* (subsequent-mention definites that refer to an antecedent with the same head noun), *bridging descriptions* (including both cases where the

antecedent denotes the same discourse entity, but using a different head, and associative bridges), and *discourse-new* cases which cannot be related to given information.

It has to be noted that besides Vieira's notion of (coreferent as opposed to associative or other) bridging definite descriptions, no usable term has been established for this class: Kunz and Hansen-Schirra (2003) simply use the term '*Is-A-relations*' (which is a misnomer since the semantic relation between anaphor and antecedent descriptions displays considerable variance beyond simple subsumption, cf. section 2.3.3), while Vieira *et al.* (2002) use the term *indirect coreference*, which clashes with the notion of *indirect anaphora* established by Schwarz (2000). (The latter term, *indirect anaphora*, is incidentally also used for coreferent bridging by Gasperin and Vieira, 2004, sowing further confusion).

1.5 Overview of the Thesis

The remaining chapters are organized in the following fashion:

Chapter 2 introduces two notions that are central to this thesis – the semantic contribution of definite descriptions, as well as the notion of coindexation in a mental model as a possible theoretical underpinning for the computational task of coreference resolution – and discusses them from a linguistic perspective.

Section 2.1 introduces some basic assumptions and clarifies some terminology that is central to this thesis. In sections 2.2 to 2.4, I review theoretical accounts of the function of definite articles, and discuss whether the generalization that definite descriptions (i.e., noun phrases with a definite article) are discourse-old (or less strongly, associated with givenness) is warranted. Finally, the notion of coreference as an equivalence relation is examined in section 2.5.

Chapter 3 introduces two fundamental ingredients to data-driven coreference resolution, namely annotated corpora and methods for quantitative evaluation. Annotation schemes for coreference resolution, as well as available referentially annotated corpora, are discussed in section 3.1, whereas evaluation settings and common evaluation metrics, which are necessary to understand quantitative results reported in the literature, are discussed in section 3.2.

Chapter 4 reviews computational approaches to coreference resolution, covering the ingredients that are used in current state-of-the-art coreference systems (sections 4.1 and 4.2) as well as existing approaches to use semantic information in coreference resolution (section 4.3). Section 4.4 reviews the state of the art in anaphora and coreference resolution for German.

Chapter 5 offers a survey of language resources for German that can be used – either directly or indirectly – to inform the coreference resolution of definite descriptions. Lexical information including morphological analysis, but also more content-oriented resources such as gazetteers and semantic lexicons, is discussed in section 5.1. Section 5.2 presents annotated and unannotated general-domain corpora for German. Processing tools that allow to extract grammatical relations

from unannotated corpora (i.e., parsers) are discussed in section 5.3.

Chapter 6 takes a closer look at unsupervised learning methods for similarity and hyperonymy relations that can be used to resolve anaphoric definite descriptions to a non-same-head antecedent. Section 6.1 gives an introduction to distributional similarity measures and the different design choices that have to be made in implementing a distributional similarity measure. Section 6.2 describes the implementation of three such distributional similarity measures for German using automatically parsed data from the newspaper *die tageszeitung* and compound splitting to counterbalance the influence of German synthetic compounds. Pattern-based identification of instance relations such as *Monsanto–company* using data from the World Wide Web is discussed in section 6.2.3. Finally, section 6.4 presents experiments on using the above information sources to identify non-same-head antecedents for discourse-old definite descriptions, a task closely related to the resolution of definite descriptions which will be the focus of the subsequent chapter.

Chapter 7 presents a system for resolving names and definite descriptions in a coreference task. In contrast to the experiments in chapter 6, no information about the discourse-old/discourse-new status is assumed and filtering of likely discourse-new mentions is carried out by the system itself. Section 7.1 describes the general setting, and the framework used to resolve names and same-head definites. Finally, section 7.2 presents experiments using semantic features that allow the resolution of non-same-head definite descriptions (coreferent bridging).

Chapter 8 summarizes the findings of the other chapters and gives an outlook regarding future work.

Chapter 2

Coreference and Definite Descriptions

In this chapter, I will discuss linguistic aspects of definite descriptions and a possible interpretation of the coreference task in linguistic terms. In section 2.1, I will argue for a linguistic model of coreference that is based on the notion of discourse referents advocated by Karttunen (1976). Such an approach makes it possible to subsume both identity-of-reference anaphora and co-reference to known entities (using names of unique designators) under a common view of *coindexation* (or co-specification, to use the term established by Sidner, 1979).

In section 2.2, I will present the two main views on the meaning of the definite article: One is that of *uniqueness*, whereby a definite description picks out the one and only thing that matches the description from the corresponding noun phrase. The other is the notion of *familiarity*, whereby definite descriptions denote a thing that is known, or at least not completely new, to the hearer or reader. Section 2.3 discusses the constraints that the models discussed in the preceding section, and related theories, would posit on coreference resolution. Section 2.4 discusses the role of salience/givenness and so-called *accessibility* or *givenness hierarchies* in the use of definite description.

Finally, section 2.5 comes back to the idea of coreference as coindexation and discusses cases where basic assumptions – namely, the role of coreference as an equivalence relation – do not seem to hold. After reviewing several examples of such cases – vague entities, blended spaces, or incompatible precisification of a mention that is deliberately vague – I propose an explanation of the acceptability of antecedent ambiguity through the notion of dot objects.

Section 2.6 provides a summary of the chapter.

2.1 A linguistic perspective on coreference

In chapter 1, I have pointed out that coreference and anaphora are distinct concepts, with the notion of anaphora having a primarily linguistic background and corefer-

ence (in the sense of reconstructing sets of mentions that refer to the same entity) being primarily motivated as a backdrop to a computational or annotation task.

A linguistic understanding of the coreference task can help to avoid inconsistencies in the definition of annotation guidelines (for example, van Deemter and Kibble, 2000 point out several avoidable inconsistencies in the MUC guidelines for coreference annotation, which will be examined in section 3.1.1). Such an understanding is also necessary to profit from linguistic generalizations in coreference resolution itself.

To see why a discussion about the linguistic background of coreference resolution would be necessary in the first place, let us consider the following intuition about reference and coreference:

(2.1) Mentions in a text can *refer* to real-world entities. When I think of the Pope, I have a particular person in mind that I can see on TV, and who I know more things about that are in the text. We can call two mentions *coreferent* when they point to the same real-world entity. Therefore, coreference is an equivalence relation (transitive, symmetric, and reflexive) between mentions in the text.

When researchers provide a definition for coreference, they commonly follow this idea of "external" reference:

Mitkov (2005) writes "*When the anaphor refers to an antecedent and when both have the same referent in the real world, they are termed coreferential. Therefore, coreference is the act of referring to the same referent in the real world.*".

Kibble and van Deemter (2000) cite Trask (1993) with the definition of coreference as "*the relation which obtains between two NPs (usually two NPs in a single sentence) both of which are interpreted as referring to the same extralinguistic entity. In linguistic representations, coreference is conventionally denoted by coindexing*". Based on this definition, Kibble and van Deemter argue that NPs in hypothetical contexts are not referring (and consequently cannot take part in a coreference relation).

This 'external reference' view corresponds to the view of Russell (1905), who takes reference as unique reference, and would allow unique descriptions such as "*Ridley Scott*" and "*The author of Waverley*" to co-refer. Such a view of reference independent of text or discourse can be used to extend the notion of coreference past a single discourse (so-called *cross-document coreference*), but, as we will see in the rest of this section, is not always adequate for linguistic models of coreference, especially when the text references a fictional or hypothetical situation.

I will argue for the following view, which is a more indirect one since it assumes a shared representation that does not necessarily have a direct correspondence to the real world.

Because it decouples the linguistic problem of building a representation for a discourse from the philosophical problem of extralinguistic reference, such a model

is potentially more useful for linguistic analysis, and subsumes both Russellian unique reference and identity-of-reference anaphora:

(2.1') Mentions in a text are related to entities in a mental model. We call two mentions *coreferent* when a competent reader interprets them as pertaining to the same entity in the discourse model. In the common case, coreference is an equivalence relation between mentions in the text, but complications in mental models can lead to violations of equivalence assumptions in rare cases.

The notion of a **discourse model** containing the entities mentioned in the previous discourse has been very influential in the linguistics of discourse. It has led to developments such as the Discourse Representation Theory (DRT) of Kamp (1981), which relates the meaning composition of single sentences to the incremental construction of a logical form for a whole discourse, but also to Fauconnier's (1984) notion of mental spaces as a more pervasive construct of thought.[1]

To see where the idea of a discourse model helps, let us consider the following example (2.2), due to Heim (1982):

(2.2) When [$_1$ a dog] barks at [$_2$ a cat], [$_2$ the cat] (always) meows.

It is clear that *the cat that is barked at* is the same as *the cat that meows*. Yet both '*a dog*' and '*a cat*' are bound by the quantifier introduced by the generic *when*, and neither mention refers to any external-world entities.

Kamp's DRT and Heim's model of file cards use the idea of *discourse referents*, due to Karttunen (1976). They each formalize the idea of discourse referents in a model that caters to (identity) anaphoric reference as well as quantification and the interaction between the two.

Karttunen introduces the idea of discourse referents by saying that an indefinite noun phrase "establishes a 'discourse referent' whenever it justifies the occurrence of a coreferential pronoun or a definite noun phrase later in the text." Karttunen uses coreference in the sense used in anaphoric binding and asserts that it "*can be studied independently of any general theory of extralinguistic reference*."

To see why it is useful to have a semantic theory that is independent from a theory of extralinguistic reference and how this is more appropriate than a purely textual account, let us look at an example from Karttunen (1976), repeated here as (2.3):

(2.3) a. Bill saw [a unicorn]. [The unicorn] had a gold mane.
b. Bill didn't see [a unicorn]. *[The unicorn] had a gold mane.

[1] Jackendoff (2002, pp. 294ff) presents a model closely related to Kamp's DRT for which he claims cognitive adequacy. Cognitive psychologists such as Sanford and Moxey (1999) or Kaup *et al.* (1999) also advocate models that track entities ("*tokens*") as well as information about these entities, but put a greater emphasis on situation-specific inferences and the connection to modality-specific representations, whereas they de-emphasize the role of quantification.

In both cases, a purely superficial account could predict that "*the unicorn*" can take "*a unicorn*" as its antecedent due to lexical substitution. For a theory of extralinguistic reference, both examples would be ontologically problematic (there is no real-world referent for a unicorn). Using Kamp's or Heim's formulations, we would be able to explain the difference by saying that in (2.3-b), the variable for "*a unicorn*" is no longer accessible because it is bound under the negation, whereas the discourse referent is accessible to further reference in (2.3-a).

2.2 Discourse-old and Definite

Unlike personal pronouns, definite descriptions are anaphoric only some of the time. To perform coreference resolution for definite descriptions in adequate quality, therefore, an accurate understanding is needed of when and how these definite descriptions have to be resolved.

Even though definite and indefinite marking is not language-universal (for example, Japanese and many Slavic languages such as Russian and Czech do have demonstrative articles, but do not have a definite article or other markers that could be seen to reliably differentiate discourse-old or unique and discourse-new entities), the definite article has the same core function in many European languages (including English and German), and is generally seen as a feature of anaphoric non-pronominal noun phrases. In the other direction – from definiteness to anaphoricity – however, the situation seems to be slightly more complicated.

A pre-theoretic notion would be that the definite article signals that a mention is somehow '*specific and identifiable*' (Halliday and Hasan, 1976). The more precise nature of this identifiability is usually explained in one of two ways:

One way is to explain the identifiability by saying that content of the definite description is specific enough that it picks out a **unique** entity. This idea dates back at least to the times of Russell (1905) and Frege (1879), and straightforwardly explains unique descriptions such as *the King of France*, but needs additional contraptions to provide a satisfying account for anaphoric uses of the definite article.

The second way to explain the identifiability is to assert that the hearer is already **familiar** with the mentioned entity, either through the previous discourse or the surrounding situation. This view was formulated by Christophersen (1939), and was adopted in different versions by Heim (1982), and by researchers such as Ariel (1990) who look at correlations between givenness/familiarity of an entity and the form of referring expression that is chosen to mention it; familiarity approaches have additional explaining to do with respect to discourse-new, especially hearer-new uses of definites (such as *the man I met yesterday*, or *the wheel* as an associative anaphor to *the car*).

In the following, I will present both the uniqueness and the familiarity analysis in their original form, but also subsequent versions of these approaches that take better care of specific problems in these accounts.

2.2.1 Definiteness as Uniqueness

Russell (1905) proposed an analysis of natural language expressions purely in terms of denoting sets (*contra* Frege (1879), who argued that both meaning and denotation were necessary parts of an analysis for natural language expressions), and was influential in laying the foundations for a broader programme to express the meaning of natural language in logical forms (whose contribution becomes clearer in more complete later works such as Montague, 1973).

Russell's analysis of the meaning of definite noun phrases is of interest here since it was the standard analysis for definites within formal semantics for a relatively long time, and some researchers advocate approaches based on Russell's definiteness-as-uniqueness analysis (which are generally referred to as 'neo-Russellian').

In Russell's model (which I will present here in first-order predicate logic notation for reasons of convenience), a natural language sentence such as "a dog sleeps" is mapped to a logical form such as $\exists x : \text{dog}(x) \wedge \text{sleep}(x)$.

Determiners correspond to quantifiers that relate the denotation of subject and predicate, which can be represented as unary predicates. For example a clause such as "*all N VP*" would be represented as

$$\forall x : N(x) \rightarrow VP(x)$$

and similarly, "*a(n) N VP*" by $\exists x : N(x) \wedge VP(x)$.

For definite descriptions, he proposes a *uniqueness* account: a definite description such as "*the moon*" or "*the bunny*" denotes the unique member of this category in the universe of discourse, and the corresponding logical form to "*the N VP*",

$$\exists x : N(x) \wedge VP(x) \wedge \forall x' : (N(x') \rightarrow x' = x)$$

asserts the *VP*ing of the unique representative of *N*.

One problem that is closely tied to Russell's formulation is that, according to this account,

(2.4) a. The king of France is bald.
b. The king of France is not bald.

are both false (assuming there is no king of France), but neither is special or anomalous in any way, conflicting with our intuition that between one normal sentence and its (natural language) negation, exactly one should be true. Strawson (1950) solves this puzzle by saying that it is necessary to draw a distinction between (i) making a unique reference and (ii) asserting that there is exactly one individual with a certain property, and that such expressions used to refer to a particular person or thing follow different rules. (The unique existence of the king of France would then be something that is now called a **presupposition**, and must be true in order to either utter a given sentence or its negation, as in (2.4)).

The bigger problem is that the uniqueness criterion needs to be amended in cases such as example (2.5), adapted from Kadmon (1990):

(2.5) Once upon a time, Leif bought a chair. He got the chair at a tag sale.

In a story about the only pink armchair Leif ever bought, the definite description "*the chair*" is perfectly fine even if Leif has several other chairs. The most straightforward way to deal with this phenomenon is to assume a pragmatic restriction on the **domain of quantification** within which the definite description has to be unique. In this way, a definite description would always pick out the unique most salient entity in a model.

Kadmon (1990) points out that in some contexts, uniqueness holds not globally but in relation to a quantified variable (which is consistent with the intuition of definiteness as uniqueness, but not with definites allowing Russellian unique reference):

(2.6) Every chess set comes with a spare pawn.
It [the spare pawn] is taped to the top of the box.

Evidence from contexts that are nonreferential, or where the definite/indefinite distinction is independent of the givenness status, suggest that definiteness should be thought of as a **grammatical distinction** rather than a referential property. Such a view supports the intuition of definiteness signaling uniqueness.

For example, the definite and indefinite articles occur in syntactic contexts that are not referential, as in example (2.7):

(2.7) a. Mary called Peter *a/*the thief*.
b. Mary called Peter *the/*a greatest liar of all*.

The view of definiteness as a grammatical distinction is consistent with the finding that indefinites are inappropriate for unique predicates (see also example (2.8), due to Löbner, 1985), and that it is possible to have a contrast between *the* and *a* (ex. (2.9), cf. Abbott, 1999):

(2.8) a. The smallest prime number is 2.
b. ??A smallest prime number is 2.

(2.9) That wasn't *a* reason I left Pittsburgh, it was *the* reason.

In summary, an approach that treats definiteness as a grammatical distinction provides a good account for cases where the definite article signals uniqueness (but see section 2.2.3, below, for some puzzles within the grammatical-definiteness account). However, anaphoric uses of definites can only be explained as uniqueness if one admits a substantial role of pragmatics in the understanding of the domain of quantification over which the description is unique.

2.2.2 Definiteness as Familiarity

The familiarity explanation for definiteness starts from the insight that definite noun phrases can be used anaphorically, whereas indefinite noun phrases can only be discourse-new (Hawkins, 1978):

(2.10) a. Fred went for a ride in [$_1$ a Jaguar], and then Harry went for a ride in [$_1$ it/the Jaguar].
b. Fred went for a ride in [$_1$ a Jaguar], and then Harry went for a ride in [$_2$ one/a Jaguar].

The motivation behind the definiteness-as-familiarity hypothesis is to extend this intuitive account of definiteness to other cases in which the referent picked out is identifiable by the hearer.

Christophersen (1939), who first summarized this standpoint succinctly, distinguishes between familiarity on an "*explicit contextual basis*" (where the referent has been mentioned in the text), a "*situational basis*" (where the referent's identity is clear from the non-linguistic context, either in terms of the immediate situation or the wider contexts), or on a "*constant situational basis*" (for referents such as the devil or the sun, which are independent of the non-linguistic context). This sense of familiarity is not limited to uses where the expressions *conjure in the hearer's mind the image of the exact individual that the speaker is thinking of* (Hawkins, 1978), but can also be something that can be picked out through an unambiguous relation to something familiar to the hearer.

Hawkins develops an account of his own starting from Christophersen's approach. He summarizes his version, which also takes care of plural and mass definite noun phrases (*the Italians*, *the juice*), by saying that speakers do the following when using a definite article:

- They introduce a referent to the hearer.
- They instruct the hearer to locate the referent in some shared (i.e., mutually known or inferrable) set of objects, and
- they refer to the totality of the objects or mass within this set which satisfy the referring expression.

The last clause (maximality of the set) is needed for plural definite references such as *the wheels*. Notice that although Hawkins starts from the idea of familiarity/identifiability and then proceeds to add the maximality requirement that is needed for plural and mass definites (*the wheels* / *the water*), he ends up with something very similar to Kadmon's (1990) Neo-Russellian approach which starts from uniqueness/maximality and then requires a contextually supplied property to restrict the domain of quantification.

A purely pragmatic account of the choice of the set, as one might interpret Hawkins' or Kadmon's maximality-over-a-pragmatically-chosen-set approach, would however be at odds with the (non-)acceptability of Partee's "marble discourse" (cited from Heim, 1982, where it is attributed to Partee):

(2.11) Ten of my marbles rolled away, and I managed to find only nine of them right away. *Later, I found it/the marble under the sofa.

Heim (1982) approaches the idea of indefinites and definites as a difference in familiarity within the context of a framework for dynamic semantics: Heim's 'File Change Semantics' and Kamp's Discourse Representation Theory, as well as Dynamic Logic (Groenendijk and Stokhof, 1984), provide a mechanism for composing the meaning of sentences that is extendable to multi-sentence fragments and (coreferent) anaphora.

Dynamic Semantics starts from the work of Montague (1973) but adds mechanisms such that (i) the logical representations of multiple sentences can be merged and (ii) resolution of anaphora can be done within the logical form.

As an example, let us take the following example of a bound anaphor, due to Geach (1967):

(2.12) Every farmer who owns a donkey beats it.

We would expect a discourse representation structure (DRS) similar to the following:[2]

$$\begin{array}{|l|}\hline x\ y \\ \hline \mathrm{farmer}(x) \\ \mathrm{donkey}(y) \\ \mathrm{owns}(x,y) \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline \\ \hline \mathrm{beats}(x,y) \\ \hline \end{array}$$

DRT incorporates pronoun anaphora and their resolution by representing them as incomplete DRSs $\begin{array}{|l|}\hline u \\ \hline u = ? \\ \hline \end{array}$; resolution is performed by replacing the ? by an accessible referent. In the example, we would first introduce an anaphoric referent u for "*it*" and resolve the anaphor to y, adding the coindexation $u = y$. The resolved structure would look like this:

$$\begin{array}{|l|}\hline x\ y \\ \hline \mathrm{farmer}(x) \\ \mathrm{donkey}(y) \\ \mathrm{owns}(x,y) \\ \hline \end{array} \Rightarrow \begin{array}{|l|}\hline u \\ \hline \mathrm{beats}(x,u) \\ u = y \\ \hline \end{array}$$

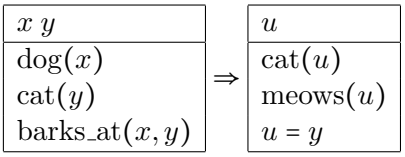
[2] For an introduction to DRT, see Kamp and Reyle (1993) or Blackburn and Bos (1999).

Heim uses a similar mechanism to account for anaphoric definite noun phrases:[3]

- An **indefinite noun phrase** (*one*, *a car*) introduces a new referential index
- a **definite noun phrase** (*it*, *the car*) must be resolved to an existing referent that entails the description of the definite noun phrase.

(This is the Heim's *Novelty-Familiarity Condition*, which we will consider more explicitly in section 2.3).

With this approach, it is possible to get a bound reading of definites; Our earlier example (2.2) – "*When a dog barks at a cat, the cat (always) meows.*" – could be represented as follows:



Dynamic semantics theories provide an *accessibility* relation that can explain the unavailability of (bound) discourse referents : the consequent part (right of ⇒) can access its own variables, plus those of the precondition part (left of ⇒), plus those of the surroundings, whereas the precondition part can only access its own variables plus those of the surroundings (i.e., the top level). At the top level, neither of the variables bound by the quantifier is accessible.

When new information is attached to the top level, all the bound variables are unaccessible. This can be used to explain why "*The dog has gray hair.*" would be odd after the discourse of (2.2).

But what about discourse-new definites such as in the following example (2.13)? For these, the requirement that the definite description be resolved to an existing more-specific discourse referent cannot be fulfilled.

(2.13) a. *Someone tells me on the phone:*
The sun is shining.
b. John read [a book about Schubert] and
wrote to [the author].

Even though a definite is used for the NPs *the sun* and *the author*, the corresponding entities would not be part of the file and the Novelty-Familiarity condition would not find an eligible existing referent. Heim explains these cases by an accommodation mechanism according to the idea in Lewis (1979); loosely speaking,

[3] Heim (1982) proposes a unified treatment for both definite pronouns (e.g., *he* or *it*, but not *one*) and noun phrases with the definite article (e.g., *the donkey*). Not all researchers agree with this solution: for example, Kamp and Reyle (1993) use a familiarity approach for definite pronouns and a uniqueness account for noun phrases with a definite article.

Lewis describes a model where the common ground (metaphorically called *game state* in his article) is updated by speaker actions, which should obey certain rules (i.e., felicity conditions). In the case of rule violations (and this is the key point of Lewis' model) the hearer updates the common ground (by adding information to the common ground) so that the action becomes felicitous.

In the case of the above example (2.13-b), the author of the book is added to the file. Heim remarks here that the information added is *not* the minimal amount necessary (as in b, adding a file card with *some* author would have sufficed for this), but the minimal amount of information that (i) makes the utterance felicitous and (ii) links any introduced new referent to an existing file card or the situation at hand. (Compare the discussion on bridging and the given-new contract in section 1.4).

2.2.3 Definites and Functional Concepts

Löbner (1985) points out that both the Russellian uniqueness analysis and the givenness account used by Heim fail to cover certain cases, and argues for an analysis of definiteness as non-ambiguity. Löbner demonstrates this difference with regard to relational concepts:

(2.14) a. The mother of one of my students
b. The front page of one of these books is missing.
c. He was the son of a poor farmer.

Being a "mother of one of my students" is not at all unique (since there probably are multiple students with distinct mothers). But since each student has only one mother, the definite article is used. Similarly, since "smallest prime number" is understood to be unique as is, "*a smallest prime number*" should have the same semantics as "*the smallest prime number*" according to Russell's approach, but the version with the indefinite article sounds definitely odd.

Relational nouns, such as *wife*, *subject*, *birth*, or *roof*, in contrast to sortal nouns such as *woman*, *adult*, *event*, cannot be replaced by a conjunction of one-place predicates, since they carry an (implicit or explicit) argument: while *wife* could be paraphrased as *married woman*, "*John's wife*" could not be paraphrased as the distinctly odd "*John's married woman*".

Distinguishing several classes of functional/relational concepts according to whether a predicate additionally depends on the current situation and/or another object, and if the relation is a functional (i.e., left-unique) relation or not, Löbner proposes to distinguish *semantic definites* and *pragmatic definites*, according to the constraints that lead to their definiteness.

A **semantic definite** is an expression that typically has a functional dependency on the situation and/or another object, such as *the moon* or *the wife (of John)*. Löbner notes that *necessary* uniqueness is not required, such as in public institutions where existence and uniqueness in an arbitrary location is just presupposed

(*the post office*, *the station*, *the toilet*, *the kitchen*). In configurational uses such as (2.14-c), *actual* uniqueness is not the case either, but instead definiteness is due to the functional concept in the head of the NP.

Pragmatic definites, on the other hand, have sortal or non-functional head nouns, and depend on the particular situation (or discourse context) for unambiguous interpretation, subsuming *endophoric* definites, where a disambiguating attribute allows to interpret the NP unambiguously, and *anaphoric* definites, which refer back to an object previously introduced.

Löbner claims that the distinction between semantic and pragmatic definites is not only relevant for discourse, but also have syntactic consequences, such as differences in cliticization to a preposition for German between (configurational) semantic definites and pragmatic definites.

(2.15) a. Er muß *in das/ins Krankenhaus.
*He has to go to (*the) hospital.*
b. Er muß wieder in das/*ins Krankenhaus zurück, aus dem er schon entlassen war.
*He must go back to *(the) hospital from which he had already been discharged.*

Poesio (1994) formalizes the sortal/functional distinction in an approach sympathetic to Heim's familiarity-based account by saying that non-relational, semantic definites are handled in the way suggested by Heim's Familiarity Condition, by linking the NP to a presupposed index:

$$\textit{the student} \rightsquigarrow [\text{student}_S(x) \wedge x = D]$$

whereas the relational interpretation (which allows configurational definites) carries a presupposition of the *argument* of the relational noun:

$$\textit{the student} \rightsquigarrow [\text{student}_S(x, D)]$$

Poesio motivates his analysis by saying that semantic definites can occur in contexts which only allow *weak* NPs, if their relation argument is indefinite, as in (2.16-c), but are not allowed when it is definite (2.16-d), whereas sortal (i.e., pragmatic) definites cannot occur in such contexts, as shown in (2.16-e):

(2.16) a. There is a student in the garden.
b. *There is {him/John/the student/every student in the garden}.
c. ?There is the student of a linguist in the garden.
d. *There is the student of {Chomsky/every linguist/the linguist} in the garden.
e. *There is the student with a brown jacket in the garden.

Note that (non-specific) uses of definite non-unique relational/functional nouns are acceptable while for sortal definites, they are not:

(2.17) a. I don't want to talk to the student of a linguist.
b. ??I don't want to steal the book of a library.

2.3 Constraints on Antecedence

Considering the accounts above, what can we say about useful behaviors or resolution strategies in a coreference resolver for definite descriptions?

In particular, what are **eligible antecedents** for a definite description? And, if a definite description has a possible antecedent, can it still be **discourse-new**?

The insights of Karttunen (1976), which have provided fundamental insights for later dynamic semantics theories, are hard constraints on possible anaphoric relations (coreferent or non-coreferent) which cannot be overridden by pragmatic preferences.

As an example, negative and all-quantified sentences usually do not allow further reference to the variables bound under the quantifier:

(2.18) Bill doesn't have a car. *It is black.

Karttunen mentions several additional cases where anaphoric reference isn't possible, such as modal verbs, implicatives and with a negative implication, and non-factives:

(2.19) a. You must write *a letter* to your parents.
*They are expecting *the letter*.
b. John didn't manage to find *an apartment*.
**The apartment* has a balcony.
c. I doubt that Mary has *a car*.
*Bill has seen *it*.

Byron and Gegg-Harrison (2004) found that a filter that uses these insights to remove nonreferring noun phrases (in the predicate of a copular verb phrase) as well as inaccessible ones (indefinite noun phrases in a modal or negated verb phrase) removed 12% of all candidate noun phrases, but only results in an absolute improvement of 0.4% in state-of-the-art pronoun resolution because these algorithms already prefer antecedent candidates in different positions. In other words, such a filter is useful, but the restriction would be very unlikely to be the only (or even the most important) factor in the resolution of definite descriptions.

More interesting for the treatment of definite descriptions, however, are predictions from the familiarity approach to definites, in particular the claim made by Heim (1982, p. 370) in her *Extended Novelty-Familiarity condition* about the possible relations between a mention NP_i (in a sentence designated as ϕ) and the accessible context (the file built from the previous discourse, designated as F):

> For ϕ to be felicitous w.r.t. F, it is required for every NP_i in ϕ that:
>
> (i) if NP_i is [-definite], then $i \notin \text{Dom}(F)$;

(ii) if NP_i is [+definite], then
  (a) $i \in \text{Dom}(F)$, and
  (b) if NP_i is a formula, F entails NP_i.

To put this in DRT terms, the antecedent must be accessible, and every proposition that is part of the anaphoric definite must be implied by the accessible propositions about the antecedent. In short, the definite description must be logically **subsumed** by the contents of its antecedent.

In this fashion, (2.20-a) and similarly, (2.20-b) are acceptable, but the coreferent reading for (2.20-c) would be non-feliciteous, even if we assumed that all relevant donkeys are seven years old, gray haired, and male.[4]

(2.20) a. Every farmer who owns [$_1$ a donkey] beats [$_1$ it].
  b. Every farmer who owns [$_1$ a donkey] beats [$_1$ the donkey].
  c. ??Every farmer who owns [$_1$ a donkey] beats [$_1$ the seven year old gray-haired jack].

Following Löbner's account, we can also see that functional and relational concepts will always be realized using a definite noun phrase, just as superlatives are, regardless of discourse-new or discourse-old status.

2.3.1 Presupposition as Anaphora

Another perspective on the resolution (or accomodation) of definite noun phrase anaphora comes from van der Sandt (1992). Van der Sandt proposes a common framework for anaphora resolution and the projection of presuppositions within DRT that is an attractive solution for the problem of definite descriptions since they sometimes behave as anaphors (i.e., more pronoun-like), but also sometimes do not need to be discourse-new. Just as Heim's account, the presupposition-as-anaphora approach is inspired by the familiarity explanation of definiteness, but it makes different predictions to Heim's for cases where a more specific antecedent is available (i.e., the anaphoric definite description's content is subsumed by that of the antecedent).

In sentences like (2.21), the mention "*Jack's children*" **presupposes** that Jack has children: claiming that "*Jack's children are not bald*" would still be asserting that Jack has children. However in (2.21-b) and (2.21-c), the presupposition is bound to the conditional (in b) or disjunction (in c).

(2.21) a. Jack has children and *all of Jack's children* are bald.
  b. If Jack has children, then *all of Jack's children* are bald.
  c. Either Jack has no children or *all of Jack's children* are bald.

[4] A jack is a male donkey.

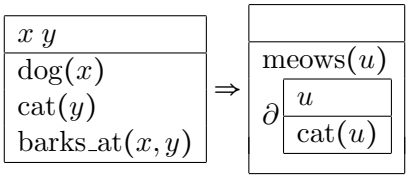
Van der Sandt claims that the projection of presupposition works similarly as the resolution of anaphora, which is also subject to binding in conditionals and quantified expressions.

The advantage of such a *representational* approach, which operates on a representation of the meaning to explain the effects of presuppositions instead of relying on logical mechanisms that result in the presuppositions falling out as a side-effect. The difference is very visible in cases such as (2.22), where an antecedent *logically* entails the information induced by the trigger (i.e., *John has grandchildren* ⇒ *John has children*), but a presuppositional reading is nonetheless preferred:

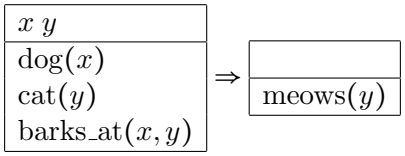
(2.22) If John has grandchildren, *his children* must be happy.

An approach similar to van der Sandt's also holds an advantage over Heim's in that her felicity conditions (essentially, the parts of her *extended novelty-familiarity condition*) are separated into a set of (hard) well-formedness conditions which have to be fulfilled by the proposed resolution, and a set of preferences, which are pragmatically determined. In the view of presupposition-as-anaphora, the difference between definite descriptions and indefinite descriptions would be that definite descriptions presuppose their content (i.e., referential index and semantic content), whereas indefinite descriptions assert it.

Formally, the example in (2.2) would yield a presupposition for the contents of "*the cat*", marked as a presupposed sub-DRS which is marked by a ∂.



The information in the presupposition can then either be bound (unified with existing information in an accessible DRS) or accommodated (added to an accessible DRS). In our case, the additional information is bound in the conditional DRS (by matching $\mathrm{cat}(x)$ and $\mathrm{cat}(u)$) which gets us the desired result:



The difference in predictions between van der Sandt's approach and Heim's Familiarity condition can be seen in cases of less-specific anaphora such as (2.23), where van der Sandt's approach predicts an ambiguity between a binding (i.e., resolution to earlier mention) and accomodation reading in cases of less-specific anaphora whereas Heim's proposal would only allow the first alternative:[5]

[5] The first example is from van der Sandt (1992); The second example is due to Krahmer and van Deemter (1998).

(2.23) a. If John has an oriental girlfriend, his girlfriend won't be happy.
b. If John buys a couple of Siamese cats, the pets won't be happy.

According to Heim's Familiarity Condition, the descriptions *his girlfriend* and *the pets* should be unambigous candidates for binding, i.e., the readings of (2.24) for the sentences in (2.23) should be non-felicitous (just as "*If John has a girlfriend, his girlfriend won't be happy*" would have only one felicitous interpretation where both *girlfriend* NPs are coindexed), but the readings in (2.24) are perfectly acceptable.

(2.24) a. John has [$_1$ a girlfriend], and if he has [$_2$ an oriental girlfriend] (as well), [$_1$ she] won't be happy.
b. John has [$_1$ pets], and if he buys [$_2$ a couple of Siamese cats], [$_1$ the pets] won't be happy.

Van der Sandt proposes that the acceptability of resolution alternatives is purely dependent on pragmatic preferences and proposes the following hard constraints based on the Gricean maxims of quality (truthfulness) and quantity (informativity):

(i) The resolution is *informative* in that the meaning assigned to the new utterance is not entailed by the meaning of the context.

(ii) The resolution is logically *consistent* (i.e., does not result in a contradictory logical form).

(iii) Informativeness and consistency must carry over to the information established in subordinate (and superordinate) binding domains.

Clause (iii) is meant to exclude cases such as the examples of (2.25), where redundancy across subordinate binding domains occurs:

(2.25) a. John has a dog. If he has a dog, he has a cat.
b. John has a dog. If he has a cat, he has no dog.
c. John has no dog. Either he has a dog or he has a cat.

Bos *et al.* (1995) propose an extension of van der Sandt's approach to cases of associative anaphora by suggesting that in cases such as (2.26), the qualia structure of the antecedent (in this case, the bar, which is supposed to have a barkeeper) is represented as a logical form in the lexicon and is inserted to introduce a referential index for the (presupposed) barkeeper:

(2.26) When I go to a bar, the barkeeper always throws me out.

2.3.2 Interaction with Discourse Relations

Asher and Lascarides (1998) present an account of definite description resolution within SDRT. The account is based on the uniqueness approach, but restricts the

quantifying domain by using a *bridging relation* B and an anchor u, which are anaphoric and have to be further specified by the context; the semantics for a definite article would then be roughly as follows:

the N VP =

$x\ B\ u$
$N(x)$
$VP(x)$
$B(x,u)$
B =?
u =?
$\begin{array}{\|l\|} \hline x' \\ \hline N(x') \\ B(x',u) \\ \hline \end{array}$ $\Rightarrow$ $\begin{array}{\|l\|} \hline \\ \hline x = x' \\ \hline \end{array}$

In this case, the B and u variables are not anaphoric in the sense discussed earlier, but are to be supplied by the context in a process that also establishes a discourse relation. (Note that the uniqueness criterion – formulated as a DRT condition $[x'|N(x'), B(x',u)] \Rightarrow [|x = x']$ – could be dropped if we required B to be a functional relation, i.e., assume that for each u, there is always only one x such that $B(x,u)$ holds; This would account for non-unique definites such as *the bathroom*).

Additional constraints then spell out a common process that jointly determines the discourse relation and the mapping of the bridge relation B and the antecedent u. In particular, Asher and Lascarides posit that an identity relation for B is preferred in exactly those cases where it is possible to assign a discourse relation, excluding a preference for identity because of a violation of temporal coherence in the *Narration* relation in (2.27):

(2.27) a. Boggs stood calmly by as Ryan struck out the hitter with [$_1$ a 95-mph pitch],
b. then he stepped up to the plate and
c. he hit [$_2$ the pitch] out of the park.

Cases where an accessible (same-head) antecedent is not chosen are usually a problem in accounts for definite descriptions, e.g. van der Sandt's preference of binding over accommodation would predict that "*the pitch*" in (2.27) would be bound to its antecedent.

2.3.3 Overspecific Noun Phrases

Already Hawkins (1978) remarks that, having introduced a person as "*the professor*", a more-specific or incomparable description such as "*the anthropologist*" is not an acceptable term for referring to that person if our hearer does not know about

the professor being an anthropologist.[6]

Hawkins' claim is compatible with the predictions made by Heim (1982) and Bos *et al.* (1995): Heim and Bos assume that subsequent mentions use only discourse-old information. Such a property would be very helpful since it would make it easier to distinguish between discourse-new definites and ones that are anaphoric. This assumption clearly does not hold for names, as it is perfectly acceptable to add new information to them:[7]

(2.28) a. Interesse [der BLG]
Interest of [the BLG]

b. Der Autoimporteur Egerland wollte auf jeden Fall selbst an eigener Kaje abladen, um [die teurere Bremer Lagerhaus-Gesellschaft (BLG)] zu umgehen.
The car import firm Egerland absolutely wanted to discharge on its own quay to avoid [the more expensive Bremer Lagerhaus-Gesellschaft (BLG)].

Since the name is a unique description, additional modification is always seen as non-restrictive. But even for common nouns, new information is often introduced by modifiers in subsequent mentions:[8]

(2.29) a. Bei einem Feuer in [einem Wohnhaus] in Bremen hat in der Nacht zum Donnerstag ein Mensch eine Rauchvergiftung erlitten.
In a fire in [a residential house] in Bremen, a man has suffered smoke poisoning on the night to Thursday.

b. Das Feuer hatte sich gegen 0.40 Uhr im dritten Stock [des viergeschossigen Hauses] entzündet.
The fire had started at about 0:40 in the third floor of [the four-storey building].

(2.30) a. Endlich nun kam [der Frankfurter "Guru" Berthold Kilian] nach Bremen, um über das aktuell diskutierte Thema Co-Abhängigkeit zu sprechen.
Finally, [the Frankfurt "Guru" Berthold Kilian] came to Bremen to talk about the currently discussed topic of co-dependency.

b. "Was Co-Abhängige tun, nannte man früher verwöhnen", erzählt [der ältere Mann mit raspelkurz-geschorenem Haar] (. . .)
"What co-dependents do has formerly been called pampering", says [the middle-aged man with the shortly clipped hair] (. . .)

In the examples, new properties such as being four-storey, or having shortly clipped

[6] As noted by Hawkins himself as well as Clark (1975), certain predicates such as *idiot* or *bastard*, called epithets, are linguistically acceptable in a more-specific or incomparable description. However, the content of such insults is clearly recognizable as hearer-new and therefore not infelicitous.

[7] TüBa-D/Z,sentences 61,68

[8] TüBa-D/Z, sentences 182,184 and sentences 210,211

hair, are included in the mention without any syntactic hints that they represent discourse-new information.

Charniak (1972) mentions that such cases of over-specific noun phrases do occur in the texts he uses (children's stories, which would be less prone to the journalistic habit of stuffing a maximum of information into phrase modifiers). He comes to the conclusion that the best way to deal with those over-specified NPs without losing the selective power of modifiers would be to detect those over-specified noun phrases and to choose the object (i.e., discourse referent) which satisfies the largest number of attributes mentioned in the NP. This would solve examples such as the following, mentioned by Charniak as example 6.22:

(2.31) Jack and Bill went outside to fly a kite. After some trouble they managed to get it to fly. Suddenly the string broke and Jack said, "My new kite is flying away."

Charniak points out that this strategy would yield undesirable results in the case of the last clause being:

(2.32) "That is OK", said Jack. "My new kite is at home."

noting that the new information ("*My new* kite") serves to mark the NP as discourse-new, as the NP would be read as discourse-old (and the reader be confused) in the absence of such modifiers:

(2.33) "That is OK", said Jack. "My kite is at home."

Charniak assumes that the distinction between discourse-new and over-specific discourse-old noun phrases has to be made using knowledge about this type of events.

Zeevat (1992), discussing cases in which the assumption that subsequent mentions do not contain new information does not hold, points to cases like example (2.34), and advocates that resolution can "*add some not implausible material and the possibility of bridging to high focus elements. The conditions under which these resolutions are not very sharply demarcated but nevertheless quite restrictive.*"

(2.34) A man died in a car crash yesterday. The Amsterdam father of four was found to have been drinking.

Zeevat argues for a process that would try resolution, also allowing bridging and adding material at the site of the antecedent. Partial accommodation – including an assumption that what information can or cannot be accommodated is subject to constraints that are external to presupposition projection proper – is an adequate explanation for the acceptability of examples like (2.34), and also correctly predicts the awkwardness of a bridge for "*his children*" using "*his grandchildren*" in van der Sandt's example (2.22). However, it means that mere **compatibility** would be the only necessary relation between a definite description and its antecedent that

could be guaranteed, unless a more specific account of overspecific definites were available.

Given Hawkins' example of *the professor-the anthropologist*, and Zeevat's restriction that the material be a plausible addition and that the antecedent must be a high-focus element, such an account does not seem completely unlikely, but is missing at this point.

2.4 Cognitive status and the choice of NP form

In contrast to the previous sections, which assumed the perspective of describing acceptable resolutions for a given definite descriptions, it is also possible to ask about the conditions under which a speaker would produce one linguistic form rather than another for a given entity (say, *he* versus *that man* versus *the man* or *the man with the large coat*).

Such an account would require two components: One component is needed to predict the salience, or *cognitive status* of a referent from the previous text. Another component would be needed to explain the choice of linguistic form based on the degree of accessibility. Different researchers in the field use different terms to express the idea of cognitive status: Ariel (1990) uses '*accessibility*', and Gundel *et al.* (1993) use the term '*givenness*', neither of which is compatible with the sense in which they are used in the other sections of this thesis.[9]

Researchers such as Ariel (1990) and Gundel *et al.* (1993) order referring expressions by what Ariel calls an **accessibility hierarchy**, meaning that expressions further up in the hierarchy tend to be realized for mentions of entities that are more accessible to the hearer. (The idea of an accessibility hierarchy could therefore be seen as a generalization of Christophersen's definiteness-as-familiarity intuition).

In the domain of definite descriptions, Ariel (2001) makes predictions for realizing entities as long definite descriptions (i.e., uniquely referring ones) versus short definite descriptions based on a difference in accessibility, and how different factors play into this difference of accessibility - for example, she claims frame-induced entities (e.g. *the waiter* in a restaurant) to be more accessible than inferable entities which are not salient or necessary in that frame (e.g., an umbrella in a restaurant, which would be realized as an anchored description such as *Maya's umbrella*). Ariel cites results where full names and long (i.e., uniquely specifying) definite descriptions are interpreted less easily than partial names or short definite descriptions when referring to a highly accessible entity (Almor, 1999).

Hierarchies such as the ones posited by Ariel (1990) or Gundel *et al.* (1993) are generally highly attractive, since in the best case they summarize several lin-

[9] Accessibility as used by Ariel is not the sense of accessibility of a referent within a dynamic semantics model as used in the previous sections, but rather the sense of salience or ease of accommodation. This confusion is due to the nonexistant interaction between the fields of dynamic semantics on one hand and cognitive linguistics on the other. Salience-based approaches to definite descriptions exist in the dynamic semantics literature (von Heusinger, 1997; Krahmer, 1998), but the rather involved mechanisms on which they are based make them largely unattractive.

guistically valid generalizations into one graded property. On the other hand, such hierarchies should not be understood as reflecting the totality of linguistically valid generalizations. Ariel points out that there is a division of labor between formal criteria (based on the linguistic category and the actual linguistic content) and accessibility requirements.

Gundel, Hedberg, and Zacharsky (1993) propose a theory that translates linguistic criteria directly to cognitive status, in the form of a hierarchy of six cognitive statuses[*sic*] relevant to the form of referring expressions:

in focus	> activated	> familiar	> uniquely identifiable	> referential	> type identifiable
{*it*}	{ *that* / *this* / *this* N }	{*that* N}	{*the* N}	{indefinite *this* N}	{*a* N}

Gundel *et al.* explain this hierarchy as a series of conditions *necessary* for the use of a certain linguistic form which imply all the conditions lower on the hierarchy. Use of lower-givenness forms is less likely due to *conversational implicature*: the use of one form normally implies that no higher hierarchy level would be appropriate, and a lower-than-usual level on the hierarchy would sound odd in most contexts.

However, Gundel *et al.*'s approach was criticized as being too simplistic by Reboul (1997) and Ariel (2001) since it makes no attempt to separate salience and other linguistic criteria.

Reboul (1997) criticizes the idea of an accessibility hierarchy determining the linguistic form as being overly simplistic. Reboul argues that the idea of an accessibility hierarchy is based on an argument that, for a pair of referring expressions, starts from the following premises

- Two different linguistic forms indicate differing degrees of accessibility.
- These two different linguistic forms transmit exactly the same information.

concludes that "the difference in accessibility is linked arbitrarily to linguistic form". This, Reboul argues, means to essentially ignore any non-truth-conditional contribution of linguistic forms, and argues that one should view the relation of linguistic form on one hand and salience/accessibility on the other hand as an epiphenomenon of reference, instead of trying to analyse any form of reference as anaphora. More specifically, she points to indexicals (you, me) and near and far demonstratives where use is clearly governed by conditions that are not covered by an 'accessibility' criterion.

Animacy and Individuation A more empirically based counterargument to Gundel *et al.*'s idea of a direct link between linguistic properties and cognitive status has been put forward by both Fraurud (1996) and Ariel (2001): animate and

inanimate entities behave differently in a highly salient context, with inanimate entities being realized preferentially as a short definite description rather than a pronoun.

Fraurud notes that even though companies and other named entities also have names, normally pronouns are preferred for persons (unless a pejorative effect is intended), while a basic-level definite description is usually preferred for non-animate entities:

(2.35) a. Ollo-Food has been very successful in the last five years.
The company has now over 50.000 employees.
b. John Smith has been very successful in the last five years.
The ?{man/person/human being} has now over 500 people under him.

Fraurud also presents quantitative results on the difference between human and non-human referents: in the texts investigated by her, no non-animate pronoun had a different-sentence antecedent, pointing to a higher locality in pronominal reference to non-animate entities; moreover, human referents are more likely to be pronominalized than non-human ones, and the majority of non-animate definite noun phrases (96%) is discourse-new (i.e., the referent has not been mentioned before in the text).

While Ariel uses such findings to advocate a more holistic approach to salience that takes into account a multitude of factors (including distance, local or global topic, or the number of competing antecedents), Fraurud (1996) argues for a more differentiated account of NP form that takes into account multiple factors.

A more complex distinction that Fraurud proposes is based on the facts that we know about entities: she distinguishes *individuals*, of which more facts are known than deductible from their general type (e.g., Peter), *functionals*, which are mainly identified by the relation they have to other entities or the situation (e.g. Peter's nose, the postman), and *instances*, where nothing is known beyond the facts that generally hold for an instance of the corresponding type (e.g., a glass of wine). This description of course holds for a particular perspective, at a particular point in discourse, as her following example demonstrates (Fraurud's translation):

(2.36) . . . och 1814 blev Schubert hjälplärare i faderns skola. Det var nog sbräcken för många års militärtjanst, som kom honom att gå med på faderns önskan. Denne hoppades å sin sida, att skolrutinen skulle få sonen att lämna "konstnärsgrillerna".
. . . and in 1814 Schubert became an assistant teacher in his [lit: the] father's school. It was probably the fear of many years' military service that made him accept his [lit: the] father's will. He [lit.: that-MASC.], on his part, hoped that the school routine would make his [lit: the] son abandon the "artist whims".

In the example, the change of perspective to the father's licenses the use of the

relational NP *the son*.

Fraurud goes on to mention that a certain degree of entrenchment (or, in her words, individuation) is a necessary condition for introducing and using names, which is only done for entities which have *individual* status.

Fraurud notes that although it would be possible to add additional properties to the factors that determine salience/givenness, including her distinction between individuals, functionals and instances, such an approach would be unsatisfactory as givenness would then become the outcome of a complex interaction between whatever factors determine NP form, and the notion of salience/givenness would become 'a rather vacuous term'. Fraurud thus argues for an analysis where salience (which captures the aspects of attentional state and previous knowledge) should be considered along with the factors captured by the cognitive ontology she outlines.

2.5 Instances of Referential Ambiguity

In section 2.1, I have argued to base coreference on Karttunen's notion of discourse referents, using the idea of *coindexation* as the underlying common phenomenon underlying the textual phenomenon of (identity-of-reference) *anaphora*, where incomplete descriptions are said to point back to earlier stretches of text, as well as *unique reference*, where independent references to a common (usually named) entity – rigid designators – are tracked to assemble the information that has been mentioned about that entity throughout the discourse.

The common view of coindexation allows us to consider a superset of both phenomena as a task for annotation and automatic resolution; but it is useful to be aware that this view is not always consistent with the (psychological) reality of an author's mental model, which finds itself encoded in a written text. Thus, it is possible to find examples which run counter to intuitions that hold for (identity-of-reference) anaphora or rigid designators. For both anaphoric and non-anaphoric subsequent mentions, the previous mention (i.e., the anaphora antecedent or previous independent mention) should be **substitutable** into the context of the subsequent mention. Under the view of coindexation, we would even expect a stronger **equivalence relation** between mentions of the same discourse referent: If A_1, A_2 and A_3 are mentions of the same discourse entity $\mathcal{A}$, any sufficiently informative description for that entity should be a plausible replacement for either of the mentions A_1 to A_3.

As a concrete example, consider the discourse in (2.37):

(2.37) [$_{A1}$ Ollo-Food] has been very successful in the last five years.
[$_{A2}$ The company] has now over 50.000 employees.

If we substitute the full name from A_1 (*Ollo-Food*) into the context of A_2 we get a perfectly acceptable sentence:

(2.38) [$_{A1}$ Ollo-Food] has now over 50.000 employees.

In the common case of discrete entities (e.g., single person referents) in normal reporting discourse, the *substitution* and *equivalence* assumptions hold as expected. In other cases, however, these assumptions do *not* hold, and it is necessary to reason more explicitly about anaphoric and reference relations to ensure a consistent treatment of such problems in coreference annotation.

I will show here how such phenomena can arise when either the entities are not discrete, but have a problem of vagueness; when the entities are well-defined but there are multiple frames of reference; or, finally, when mentions evoke (and refer to) more than one entity, as can be the case in polysemy and metonymy but also in other cases.

I use specific examples to examine some of these cases where the posited criteria do not hold, creating difficulty in the annotation[10] and investigate the nature of the resulting problems, possibly making reference to specific theories of discourse for explanation.

2.5.1 Co-reference of Vague Entities

In newspaper text (as opposed to stories told by children or drunk people), it is to be expected that the text can be interpreted without major ambiguities in interpretation, such that different people reading the text will get a very similar view of the main events or propositions of said text, even when annotating that text might uncover referential ambiguities. One working hypothesis would therefore be that referential ambiguities must be attributed to the reference relation between mentions and pieces of modeled reality, as well as the identity conditions between these pieces. Both of these are non-issues with concrete referents, or referents like mountains that have a vague extension (i.e. there can be disagreement about which parts at the bottom of the mountain still belong to it) but can be individuated by their peak. In the case of vague entities without an obvious individuation criterion, the existence of two overlapping entities of the same class is not ruled out a priori, and coreference decisions can become difficult.

Consider the following sentences, taken from the TüBa-D/Z corpus:

(2.39) a. Für ein “barrierefreies Bremen” gingen deshalb gestern [$_1$ mehrere hundert behinderte Menschen] auf die Straße – und demonstrierten für “Gleichstellung statt Barrieren”.
For a “barrier-free Bremen”, [$_1$ *several hundred disabled people] went onto the streets yesterday — and demonstrated for “Equality, not Barriers”.*

b. “Warum immer wir?” fragten [$_2$ die Versammelten] deshalb auf Plakaten.
“Why always us?” [$_2$ *the congregated*] *therefore asked on the posters.*

[10] All corpus examples presented here are disagreements between annotators in the TüBa-D/Z.

If we use both descriptions in isolation to delineate an extension for the group entities that each mention refers to (taking a very literal approach of co-reference as first determining a real-world entity for each mention and then saying that they co-refer if the two entities are the same), it is intuitively clear that the person groups from mentions 1 and 2 must have a large overlap. But, seen in isolation, the real-world extensions of the two mentions do not seem to be identical, as not every demonstrator had disabilities, and neither did every one of them carry a poster with the indicated question.

On the other hand, we would like to treat the demonstrators as one entity that is described by several predications and not several distinct entities, just as we would not want to talk about multiple clouds when there is just one cloud in the sky to which several predicates apply differently on different parts.

If we treat the conditions of being disabled and of carrying posters as incidental and instead use the demonstrating as the defining property of the crowd of mentions 1 and 2, we can coerce the individual predicates of being disabled, and wanting to push for a "barrier-free Bremen", to (vague) predicates of groups by taking a majority view. That is, the article talks about a crowd of demonstrators that

- wanted to push for a "barrier-free Bremen"
- comprised (about) several hundred people
- consisted (in a significant proportion) of disabled people
- had some posters asking "Why always us?" (cumulative reading)

This model is preferable on the grounds of being *minimal*: for two distinct entities in the model, there is some information (either in the text or from background knowledge) that distinguishes them, and it is preferable to talking about several overlapping but not identical clouds.

But this also means that we have to *first* build a model (including merging the 'referential indices' of coreferent mentions) and *then* interpret it, as opposed to first taking a relation of mentions to (real-world) entities and then judging coreference on the equality of referents, which will lead to strange results. Conversely, the model-first approach that we want to advocate here weakens the idea of *substitutability* of two textual descriptions to one of compatibility of a description with some discourse entity. Unless we are ready to admit vagueness — multiple possible extensions for a group, even though the definite article suggests that the expression has a unique denotation — we would have to define this compatibility in similar terms than for substitutability, which would still be problematic.

The proposed model would have the denotations for the real-world entities to shift in the process of incremental model-building that accompanies text comprehension, something which is not intuitive but, as we will argue, must be considered a realistic assumption.

To apply the minimal model criterion, we still need to be able to discuss whether merging two entities in the model would lead to a contradiction (assuming

common-sense background assumptions), and it might be argued that allowing discourse entities to be described by vague predicates lets the judgements on them, including compatibility or mergeability, become vague, and possibly ambiguous.

To see that this is not the case, let us consider how the problem of vagueness in reference is solved in Smith and Brogaard (2001)'s supervaluationist account of reference to vague objects and predications of these objects. Smith and Brogaard posit that one can, for a vague object, give multiple precisifications relevant to a certain context – for a cloud, several cloud-shaped sets of water molecules, for a crowd, multiple sets of persons, or, for a house that has an annex built besides it, the house with or without the annex.

A statement is then judgeable and true (supertrue) iff we can instantiate every singular term with a corresponding family of aggregates and, however we select a single possibility from the family of aggregates, the statement is true.

In example (2.39), the predicates of having a certain political objective (by extension from the members to the crowd), to comprise several hundred people, to consist (in a significant proportion) of disabled people and of carrying some posters (in a cumulative sense) are all vague, and not every precisification of one predicate is a plausible precisification of another predicate. But if we use a conjunction of all predicates (a straightforward way of interpreting this would be to take the intersection of the individual precisifications), we still get a plausible set of precisifications for the whole description. Thus, we can instantiate the term that represents our referent in the model with a family of precisifications that makes the whole description judgeable and true (which is what we wanted in the first place). The set of possible precisifications for the joint descriptions would then be a subset of the precisifications for an individual mention, but the set of *preferred* precisifications could be a different one, thus explaining a shift in the (vague) denotation in the course of incremental model-building (where we first have a description only based on the first mention, which is subsequently enrichted to include the information from the second mention).

2.5.2 Reference in Blended Spaces

The account of vague extensions presented in the preceding section allows a clearer view on disagreements between a markable being discourse-new and it being discourse-old (and coreferent to an earlier mention), by looking at plausible precisifications for referents where an individuation criterion is not available.

Some other ambiguities are due to a richer structure of the mental model than what can be represented as an equivalence relation. Annotating ambiguities would not do justice to the model structure either, since some ambiguities can be related, as in the following example (2.40):[11]

(2.40) a. [$_1$ John Travolta] verklagt als Bostoner Anwalt zwei Firmen, die er für den Leukämietod von acht Kindern verantwortlich macht.

[11] Example from TüBa-D/Z, sent. 3772ff.

As a lawyer in Boston, [$_1$ *John Travolta*] *sues two businesses that he holds responsible for eight children having died of leukemia.*

b. Anfangs wittert [$_2$ der berechnende Karriereanwalt] nur die hohe Entschädigungssumme (. . .).
At first, [$_2$ *the calculating career lawyer*] *only scents the high amount of compensation (. . .).*

c. Gerichtsdrama, Umweltthriller und großes Schauspielkino, in dem [$_3$ Travolta] und sein Gegenspieler Robert Duvall zu Hochform auflaufen.
A court drama, environmental thriller and great actors' cinema, in which [$_3$ *Travolta*] *and his antagonist Robert Duvall reach top form.*

Markables 2 (the career lawyer), the lawyer figure from the story and 3 (Travolta), the actor, are not compatible, since they mention two different entities. But markable 1 mentions John Travolta in a context that obviously belongs to the career lawyer. Annotating markable 2 and 3 as ambiguous between discourse-new and coreferent to markable 1 would miss the point, since the interpretation where 2 and 3 are both discourse-new is not felicitous. In terms of the theory of mental spaces (Fauconnier, 1984), we have two spaces, one with actors and one with the story's protagonists, and a third one where the two are blended together. Based on this account, we can choose to consider the blended space for the equivalence relation that we base coreference on.

The solution of using a blended space can also be cast in terms of the *dot objects* described by Asher and Pustejovsky (2005): To explain the semantics of sentences like "The book was a huge pain to lug home and turned out to be very uninteresting", where both the physical object and the book's content are mentioned, or "Mary read the subway wall", where an inference has to be made that yields an *informational* aspect of the subway wall, they introduce compositions of primitive types; the book would then be modeled to be of the type *physical* • *informational*, and would have as aspects both the physical object as well as the information contained in it. Asher (2006) uses this to explain the felicity of sentences like "*Superman always gets more dates than Clark Kent does.*" where the usual denotation of *Superman* and *Clark Kent* would be a single individual. By introducing aspectual objects for *Superman as Superman* and *Superman as Clark Kent* (which behave differently with respect to having dates, depending on whether Superman shows up in his Superman role or as Clark Kent) that are both aspects of the (non-aspectual) person Superman.

For our purposes, we could represent the entities from in the blended space in example (2.40) as *actor* • *role* dot objects which have both an actor aspect (reach top form in a court drama) and a fictional person aspect (sue two businesses). The coreference annotation would then consider this dot object (corresponding to an entity in the blended space) as referent for the coreference chain. This solution fails when we have multiple blended spaces for different film adaptations of the same

story, where we would get multiple dot objects with the same role, but different actors, and there would not be one preferred blended space. Such texts are rare enough not to be a problem in practice, but they exist.[12]

2.5.3 Incompatible refinements to a vague description

In some cases, we find that, on a coarse level of detail, some entity is repeatedly mentioned and then taken up with an anaphoric mention, whereas on a finer level of detail, the repeated mentions are incompatible, creating a problem of referential ambiguity[13]:

(2.41) a. Die Konzepte reichen von einseitiger Einstellung der Luftangriffe (...) bis zu einem politisch-wirtschaftlichen Marshallplan für [$_1$ den gesamten Balkan].
The concepts range from a unilateral cessation of the air strikes (...) to a politico-economic Marshall plan for [$_1$ *the whole Balkan*].

b. Zu Beginn der Woche hat US-Botschafter Robert Barry ... seine Gedanken zu einer nachhaltigen zivilen Lösung für [$_2$ die Dauer-Problemregion Ex-Jugoslawien] zusammengefasst.
At the beginning of the week, US ambassador Robert Barry has (...) summarized his thoughts regarding a sustainable civil solution for [$_2$ *the permanent problem region of ex-Yugoslavia*].

c. Barry fordert ... eine Umstrukturierung der Ausbauhilfen in [$_3$ der gesamten Region].
Barry demands (...) a restructuring of the development aid measures in [$_3$ *the whole region*].

In this case, the theme is a plan by Robert Barry to stabilize the war-riddled region around ex-Yugoslavia by means of targeted financial aid. The exact region isn't specified – and probably the plan is not this detailed yet, so the region is vague and its exact limits are underspecified. But the article mentions this region first as "*the whole Balkan*", and then as "*the (...) region of ex-Yugoslavia*", which are both plausible extensions of this region but not compatible with each other. As a result, one annotator marked markable 1 as coreferent with "the whole region", while the other chose markable 2.

A precondition for this anaphoric reference to a vague entity to be felicitous would be that the reader does not notice the incompatibility between the two previ-

[12] Consider the comparison of different adaptations of the Dickens novel "Great Expectations", found at `http://www.victorianweb.org/authors/dickens/ge/filmadapt.html`: whereas in one adaptation, "Pip rips down the draperies of Satis House to let in the light of day upon the mould and decay of Satis House and release Estella", another is showing "Pip's snobbery", all while the reader is perfectly comfortable with this over-abundance of blended spaces which would make coreference annotation overly problematic.

[13] TüBa-D/Z sentences 6387,6389,6393.

ous mentions. From a production perspective, we could argue that the vague entity that the author had in mind was coerced to a more specific one (by the unavailability of more basic terms for the vague region); from a comprehension perspective, the construction-integration model of Kintsch and van Dijk (1978) would predict that some facts from the discourse get simplified in the process of understanding (there is intervening text between the example sentences), and it would be plausible that the more specific representations of "*the whole Balkan*" and "*the (. . .) region of ex-Yugoslavia*" are reduced to the underlying vague representation that we posit here.

We can formalise this by saying that the mentions evoke both an entity from a space with precise representations, and one from a space with coarser representations, yielding a composite object with a vague and a precise aspect (where the vague aspect is needed for the structuring/comprehension of discourse) – this would not help us immediately for possible solutions since the precise aspect could be referred to in a different context (consider a similar text where "*the (. . .) region of ex-Yugoslavia*" is followed by mentions of ex-Yugoslavia that elaborate on other propositions and thus don't have the vague region aspect that the first mention has). But we gain at least some explanatory adequacy by the use of this device since we can then argue about the presence of the additional aspect pertaining to a coarser frame of reference.

Positing dot objects with aspects for different levels of granularity – which would correspond to the 'local' theories of different granularity levels, as posited by Hobbs (1985) – would provide an insight into other cases of coreference with two incompatible (potential) antecedents, as in the following example from (Knees, 2006)[14]:

(2.42)
a. It was Robert Jackson's will that this should not remain an isolated case; [At that time already, he demanded that an International Criminal Court should be established].
b. He wanted to take advantage of the favourable moment as the world was shocked by the atrocities of the Nazis, the full enormity of which could only be anticipated first at [Nuremberg]. (. . .)
c. But an International Criminal Court only starts gradually to take shape today, 50 years after [that], in The Hague, where the war crimes of the former Yugoslavia are being dealt with.

Both the Nuremberg trials and Jackson's demanding are plausible antecedents for the anaphoric "after that", but they are obviously *not* compatible. We can model this by positing a coarser granularity level (since the anchoring description *50 years after that* is coarser than the granularity of the anchoring events *the Nuremberg trials* and *Jackson's demanding*) and saying that the anchoring description coerces the antecedent descriptions to a coarser granularity.

[14]Translated from the German original version containing the pronominal adverb *danach*

Knees remarks that this could be an example for the Justified Sloppiness Hypothesis brought forth in (Poesio *et al.*, 2003), which was motivated by examples like the following, which Poesio and Reyle (2001) found in their study of the TRAINS corpus:

(2.43) a. Hook up [engine E2] to [the boxcar] at Elmira.
b. And send [it] to Corning as soon as possible.
c. As soon as it arrives, [it] should be filled with oranges.

In this case, annotators disagreed whether the pronoun "it" was referring to the boxcar or the engine. Poesio and Reyle argue that the pronoun (in b) could be seen as referring either to the boxcar, the engine or the whole train (as the mereological sum of both), and argue that a theory of discourse might leave the pronoun interpretation under-specified between all three solutions. In a case like the above including (c), the under-specification gets fully specified later (in c), since only boxcars can be filled with oranges.

Poesio *et al.* account for this phenomenon with the hypothesis that ambiguity between multiple potential antecedents can occur without being rejected as being ungrammatical (which would certainly occur in the case of an ambiguity that allows for different scenario interpretations, a case where we would see repairs in dialogue or largely negative judgements in a magnitude estimation study), as long as the following conditions are met:

1. Both explicitly mentioned potential antecedents x and y are elements of an underlying mereological structure with summum $\sigma = x \oplus y$ which has been explicitly constructed (and made salient) in the dialogue.

2. The existence of this structure makes it possible to construct a p-underspecified interpretation[15] in which the anaphoric expression is interpreted as denoting an element z included in the mereological structure – i.e. part-of its summum σ:

$$[x\, y\, \sigma\, z \mid \ldots \sigma = x \oplus y,\ z \lhd^* \sigma \ldots]$$

3. All possible interpretations (x, y, z, $x \oplus y$) are equivalent for the purposes of the plan.

Assuming a mereological hierarchy like that of Reitsma and Bittner (2003)[16], the Justified Sloppiness Hypothesis provides a good explanation for examples like

[15] In analogy to lexical homonomy and polysemy, Poesio *et al.* use the term p-underspecification for cases where the underspecified form can be interpreted, in contrast to cases that they call h-underspecification, in which interpretation necessarily resolves the ambiguity present in the text.

[16] Link (1983)'s original version contained two hierarchies, one of atomic objects and plurals composed of several atomic objects, and one of unstructured lumps, connected by an extension function that maps (groups of) atoms to the lump that is their material extension. Since σ is not a plural and not a simple sum of extensions (as in "the material making up John's left hand and the Empire State Building"), we need to posit a (possibly domain-specific) mereology of functional or otherwise salient parts and wholes.

(2.42) and (2.43). However, we could not use the Justified Sloppiness Hypothesis to explain our earlier example (2.41) since the precisifications are not mereological parts of the vague region; also, the reference to mereological structure could be seen as somewhat arbitrary.

In our model, the trains example would be explained by the fact that "*send ? to Corning*" selects for a train and thus triggers the dot-introduction, leading to a *boxcar* • *train* object where only the train aspect is used. Since the other aspect is not used, we can then posit underspecification between a *boxcar* • *train* and a *engine* • *train* entity as referent for the "*it*" in (b). The mereological structure then serves to fulfill the precondition for the introduction of dot objects, but is not special otherwise.

We can formalise this extension of Poesio et al.'s Justified Sloppiness Hypothesis as the following *Generalised Sloppiness Hypothesis*:

Multiple potential antecedents can occur without being rejected as ungrammatical as long as

1. The anaphoric expression occurs as argument to a predicate $P : \beta \multimap t$, i.e. the context selects for a type β.

2. The potential antecedents $x : \alpha_1$, $y : \alpha_2$ have to be extended to a β-compatible complex type (i.e., $\alpha_1 \bullet \beta$ or $\alpha_2 \bullet \beta$).

3. In both cases, the same entity $z : \beta$ is selected as result of the dot introduction.

The notion of dot objects allows us to treat blended spaces, granularity shifts, and cases of polysemy in a uniform way, much like Mani's (1998) treatment of granularity shifts and polysemy using abstraction operators, and it allows us to treat a greater range of phenomena than with Poesio et al's original formulation.

It is also possible to predict the acceptability (or, in Poesio's terms, justification) of such examples as resulting of (i) the plausibility of a coercion from $\alpha_{1/2}$ to β as a result of the context (P) and (ii) the interference from a reading without coercion which would make the example unacceptable.

The difference between coercibility criteria and Poesio et al.'s formulation using plans would be visible in a hypothetical case where the plan is to blow up both the engine and the boxcar by joining them together and planting a bomb on them:

(2.44) a. Hook up [engine E2] to [the boxcar] at Elmira.
b. Attach a large bomb to [it] and blow it up.

For the purposes of the plan, it is not important whether the bomb is planted on the engine, the boxcar or the train (assuming that the bomb is large enough), but the predicate of 'attaching a bomb to something' does not require its argument to be a train, and no dot introduction is involved. My Generalized Sloppiness criterion would thus (correctly) predict that (2.44-b) is awkward.

2.6 Summary

In empirically oriented work in computational linguistics, it is often stated that "Coreference resolution is the process of determining whether two expressions in natural language refer to the same entity in the world" (Soon *et al.*, 2001), or that two markables are coreferential when they "refer to the same object, set, activity, etc." (Hirschman and Chinchor, 1997). While there is no denying that we do want to know what's being said about the real-world entities we know and care about, this kind of definition turns out to be too coarse due to intricacies of (linguistic and general) reference, and trying to do coreference annotation on this basis will invariably run into difficulties, as e.g. van Deemter and Kibble (2000) are quick to point out. In the course of this chapter it has become clear how coindexation in a mental model can serve as the basis for coreference annotation, allowing for a task definition that encompasses both coreference (in the stricter sense of unique descriptions) and identity-of-reference anaphora, but also where the limits of such an approach lie.

Taking the semantic contribution of the definite/indefinite distinction and proposed treatments in (non-computational) linguistics, I have pointed out useful assumptions that coreference resolution can use (regarding a compatibility, equivalence, or subsumption relation between an antecedent and the definite description we want to resolve), and also made reference to partly disillusioning results that resolution of definites may, in some cases, be as hairy and complicated as presuppositions in general, or co-dependent on discourse structure even in cases that are devoid of further complications such as associative or abstract object anaphora.

The review of sections 2.3 through 2.4 outlines several of the **central problems** for the resolution of definite descriptions: one is the question of admissible relations between an anaphoric definite description and an antecedent, as well as the related question of recognizing definite descriptions that are discourse-new (i.e., either non-anaphoric or introduced by a non-identity link).

The discussion in this chapter made clear that even short descriptions that are very likely to be incomplete (e.g., *the city*) are not guaranteed to be discourse-old, as they could be exophoric, or be accommodated through a non-coreferent relation. Since such descriptions usually have a conceptual basis – consider the idea of functional and relational definites put forward by Löbner (1985) – such *potentially* discourse-new definites can be identified using data-driven methods such as those presented in chapter 4, but the question whether they are actually discourse-old or discourse-new can only be answered with respect to an actual context, which is why the system presented in chapter 6 uses an integrated classification approach that covers both discourse-new detection and resolution.

The review of section 2.3 also uncovers several plausible constraints on the antecedents of anaphoric definite descriptions and their content: Heim's *extended novelty-familiarity condition* requires that the content of the subsequent mention be entailed by the content of the previous mentions (**subsumption**: *a dog* would

allow a subsequent mention of *the animal*, but not of *the oversized cocker spaniel*). Van der Sandt's approach of presupposition as anaphora is less constrained in that it only requires that a resolution adds new information and offers **consistency** with the existing model. However, it also predicts greater ambiguity than Heim's model in the cases where a mention could both be understood anaphorically and has a plausible way of accommodation, because the latter rules out accommodation where an acceptable anaphoric resolution is possible. Both notions used to explain the semantic behavior in the resolution of definites – the notions of *subsumption* and *consistency* – offer a different perspective on the search for antecedents from the idea of semantic similarity. Even though it is intuitively plausible, semantic similarity is rather harder to pin down from a logical perspective than subsumption or consistency.

Chapter 3

Coreference Corpora and Evaluation

This chapter discusses two issues that are central to understanding results published in the literature on coreference resolution: one is the annotation scheme of coreferentially annotated corpora (and, by extension, the definition of the coreference task that can be derived from it), discussed in section 3.1, while the other issue consists in the exact evaluation setting, and the evaluation metric used, to perform a quantitative evaluation, discussed in section 3.2.

In the 1990s, a series of events called Message Understanding Conferences (MUCs) introduced formal evalution on real-world data, by then a standard procedure in information retrieval, to the (purpose-driven) analysis of written text. In a task that was both considered useful for actual application and sufficiently nontrivial that it would showcase the faculties of then-current natural language understanding systems, *information extraction* systems had to fill a domain-specific template with information from the text.

With some progress in this field, it became increasingly clear that it is necessary to merge (and reconcile) multiple pieces of information about the same object when they are found in different sentences of the text (Hirschman, 1992; Bagga and Bierman, 1997). Motivated by these findings, a separate coreference task was spun off in the sixth MUC – again, empirically grounded using **annotated data** (a collection of real-world texts) and **formal evaluation procedures**. The annotation effort for MUC-6 (see section 3.1.1) did not limit itself to mentions that were immediately useful for the template extraction task; Instead, spans referring to any kind of entity were marked and annotated by assigning a set identifier to all mentions in a coreference chain.[1]

The availability of annotated data led to renewed interest for anaphora and

[1] An earlier effort with a slightly more ambitious scope is the IBM-sponsored UCREL corpus (Fligelstone, 1992), which included not only coreference annotation, but also ellipsis, inferred complements and abstract object anaphora. The fact that UCREL's corpus was never made available to the community at large is the most likely reason that the effort went largely unnoticed.

coreference resolution in computational linguistics. The resolution of "full noun phrases" (i.e., non-pronouns, including definite descriptions and names), had up to that point lived more of a fringe existence,[2] but was revealed to offer interesting challenges when real-world data was examined.

Involvement from linguistically interested parts of the community and subsequent critique of the annotation guidelines underlying the MUC corpus led to annotation efforts that were more well-informed linguistically than the latter (see section 3.1).

In parallel, the evaluation procedure proposed in the MUC efforts – or the question of evaluating coreference systems in general – was subject to a significant amount of discussion. Proposals that address real and perceived shortcomings of existing metrics were adopted by subsequent research. On one hand, the newly proposed metrics are less sensitive to over- or under-merging, thereby allowing for a more reliable tracking of improvements to a baseline. On the other hand, the substantial variation in the evaluation used within the existing literature makes it necessary to understand evaluation measures, and their interaction with the exact setting used in the evaluation, to assess the state of the art.

3.1 Available Corpora and their Annotation

In the last years, large referentially annotated corpora not only for English, but for a wide variety of languages have been created, including the referential layer of the TüBa-D/Z corpus. While the annotation schemes usually differ in their exact details, a majority of them is either inspired by one of the frequently used corpora for English – the MUC and ACE corpora, or by the MATE proposal for coreference annotation, which is meant to cover a greater variety of languages and text genres.

In terms of the actual encoding of the annotation, and the annotation task itself, coreference chains can either be represented in a *link-based* fashion, where anaphoric identity-of-reference links are marked (by adding a pointer to the antecedent markable), or in a *set-based* fashion, where coreference sets are represented explicitly (by adding one common identifier to all mentions of one coreference chain). Both representations are found in common corpora, but either can easily be converted into the other – in one case, by finding weakly connected components in the antecedence graph, in the other, by introducing anaphoric links to the most recent element of the same coreference sets. Hence, I will continue to use the set-based representation of coreference chains, as in (a), even though some corpora represent the same information as in (b).

(3.1) a. $[_1$ Peter] thought $[_1$ he] was dreaming.
b. $[_a$ Peter] thought $[_{b \rightarrow a}$ he] was dreaming.

[2] For example, Carter (1985) implemented the resolution of definite descriptions using heuristics proposed by Sidner (1979), but only reports results for pronoun resolution on his sample of small stories created for that purpose.

<COREF ID="1">*Edna Fribble*</COREF> and
<COREF ID="2">*Sam Morton*</COREF> addressed the meeting yesterday.

<COREF ID="3" REF="1" TYPE="IDENT" MIN="Fribble">
Ms. Fribble</COREF> discussed coreference, and
<COREF ID="4" REF="2" TYPE="IDENT" MIN="Morton">
Mr. Morton</COREF> discussed unnamed entities.

<COREF ID="1" MIN="Fribble">*Ms. Fribble*</COREF> was
<COREF ID="2" REF="1" TYPE="IDENT">*president*</COREF>
and
<COREF ID="3" REF="1" TYPE="IDENT" MIN="CEO">*CEO of Amalgamated Text Processing Inc.*</COREF>

Figure 3.1: Markup in the MUC scheme

3.1.1 The MUC scheme

The earliest larger-scale corpora to be widely used were the coreferentially annotated corpora of the 6th and 7th Message Understanding Conferences.[3] Born out of the observation that coreference resolution was a crucial component for the task of Information Extraction that was explored in the earlier MUC conferences (an observation made, among others, by McCarthy and Lehnert 1995, citing participants from earlier MUCs), the task of coreference resolution was instituted as a separate task.

As the MUC coreference task aims at coreference annotation for newspaper text, it concentrates on nominal coreference and uses a very pragmatic definition of coreference as nominal elements being coreferential whenever "they refer to the same object, set, activity etc.", without requiring that the relation is anaphoric in nature (Hirschman and Chinchor, 1997).[4]

In an SGML-based scheme, coreference relations can be annotated between nouns, noun phrases and pronouns (including possessives). Names and dates are considered opaque and parts of them (e.g., *Iowa* in *Equitable of Iowa Cos.*) are not to be annotated. Annotated markables contain a MIN attribute that makes it possible to evaluate coreference results on system-generated markables against the

[3] These corpora are available from the Linguistic Data Consortium as:
Nancy Chinchor and Beth Sundheim (2003): Message Understanding Conference (MUC) 6, LDC2003T13
Nancy Chinchor and Beth Sundheim (2001): Message Understanding Conference (MUC) 7, LDC2001T02

[4] Hirschman and Chinchor (1997) write: "*It is not a requirement that one of the markables is 'semantically dependent' on the other, or is an anaphoric phrase*". As an illustration of this idea, consider example (1.8) from chapter 1, where "*William Jefferson 'Bill' Clinton*" and "*Bill Clinton*" are coreferent without either of the mentions being anaphoric.

gold standard even when their spans do not line up exactly (for example in the case of parsing errors). Figure 3.1 shows the link-based annotation in MUC with MIN attribute.

The relations to be marked up according to the MUC scheme encompass coreference (which, as pointed out in chapter 2, should be understood as coindexation in a discourse model, and not with the Russellian sense of reference in mind), including bound anaphora, but also cases where the type of a generic noun phrase is invoked by a premodifying noun, or where a (usually non-referring) predication is syntactically linked to a noun phrase.[5]

- Evoked prenominal constructions:
  The [$_1$ aluminium] price ... [$_1$ Aluminium] is ...
- Appositive constructions:
  [$_1$ Peter Miller], [$_1$ president of Sudsy Soap Co.]
- Copula constructions:
  [$_1$ Peter Miller] is/became [$_1$ the president of Sudsy Soap Co.]

Both the inclusion of bound anaphora and the indiscriminate treatment of predicating constructions were criticized by van Deemter and Kibble (2000), who maintain that this kind of coreference definition lumps together several phenomena that should not be conflated, and that this leads to problematic cases like

(3.2) a. *The stock price* fell from *$4.02* to *$3.85*;
b. Later that day, it fell to an even lower value, at *$3.82*.

Considering only (a), the provision in the MUC guidelines – that only the latest value is to be annotated among several incompatible ones – would prescribe that *$3.85* be annotated as coreferent to *the stock price*, whereas considering all the text, it would have to be *$3.82*.

3.1.2 The MATE coreference scheme

After the MUC effort, many other annotation efforts have proposed similar guidelines, with some variation around what one could call "contentious" issues – the treatment of copula and predicates (which are not part of coreference proper, but are frequently subsumed under the notion of coreference resolution), as well as metonymy (explained below). All guidelines, however, adhere to the basic principles of marking text spans and grouping them into coreference chains either by means of antecedence links or by assigning set identifiers that are unique to each coreference chain.

[5] The treatment of premodifying bare nouns in common noun phrases (which are to be annotated) is at odds with the treatment of parts of a name (which are not to be treated as markable). The most likely reason for this is a desire to keep consistency with MUC's named entity annotation task, where names are also treated as opaque text spans.

The MATE coreference scheme (Mengel *et al.*, 2000; Poesio, 2004) is a proposal for coreference annotation that is meant to encompass a larger domain of application than the MUC coreference scheme, including other languages and different text types such as task-oriented dialogues.

In English, the MATE scheme has been used for the GNOME corpus (Poesio *et al.*, 1999), which consists of museum labels, pharmaceutical leaflets, and tutorial dialogues; in the VENEX corpus (Poesio *et al.*, 2004b), the MATE scheme was used to annotate spoken and written Italian.

To account for clitics and zero subjects, as they are frequently found in Romance languages such as Italian or Spanish, the MATE scheme allows to use existing tokens as proxy for the actual zero element or clitic. In contrast to normal markables, which are marked with the `de` tag (for "discourse entity"), such 'proxy' markables are marked with a `seg` tag.

To account for the frequent cases of deixis in task-oriented dialogues, the scheme allows for *universe entities* (i.e., a list of entities that are in front of the dialogue participants), which can be used as anchor for coreference links.

With respect to the core coreference scheme the MATE scheme introduces some improvements over MUC's notions: anaphoric links to quantified noun phrases (such as "*every man*") receive the label *bound*, bare nouns are not markables.

Predicate noun phrases such as those in copula clauses are also treated differently: cases such as the example (3.2) are marked with a *function-value* relation, whereas *predicate nominals* (such as '*a policeman*' in '*John is a policeman*') are not annotated. Poesio (2004) reports that excluding these noun phrases from the annotation makes automatic markable extraction more difficult; for the GNOME corpus, this was solved by introducing a new attribute which is then used to distinguish referring (in the strong sense) NPs, which receive a value of `term`, quantifiers (`quant`) and predicative nominals `pred`.

The MATE proposal also includes an *extended scheme* that aims to capture non-identity anaphoric relations such as the bridging relations identified by Passonneau (1996) and Vieira and Teufel (1997):

- *set relations* such as *membership* and *subset*
- *possession* and *part-of* relations
- *instantiation of concepts*, where a category-denoting term is linked to a specific instantiation:

  (3.3) We need [$_1$ oranges].
  There are [$_{2\rightarrow1}$ some] in Corning.

- *event relations* including causation, such as *explosion – noise*.

  The proposal also mentions annotation of clausal antecedents to event-related bridging relations, such as in the following example:

(3.4) ⟨$_1$ Muslims from all over the world were taught gun-making and guerilla warfare in Afghanistan⟩.
[$_{2\rightarrow1}$The instructors] were members of some of the most radical Islamic militant groups in the region.

Poesio (2004) mentions that the actual inventory used in the GNOME corpus was much more limited, as many of the bridging relations, including attribute relations (such as *car – price*) and situational associations (such as *restaurant – waiter*) were difficult to annotate and showed low annotator agreement.

3.1.3 The ACE coreference task

In the context of information extraction, coreference resolution is most important for a small number of semantic classes that are important to the domain at hand; indeed, many early machine learning approaches such as those of McCarthy and Lehnert (1995) and Aone and Bennett (1995) only concerned themselves with organizations and persons. One potential benefit of narrowly focusing on a small number of (presumably) well-behaved semantic classes is that identity or non-identity is usually straightforward to determine, whereas it may be very difficult to decide it for abstract or vague objects. To improve the consistency of the annotation, the coreference task in the ACE evaluation[6] limits the consideration to coreference links between persons, organizations, geopolitical entities (i.e., countries and regions with local government), locations, vehicles and weapons.

A problem in coreference annotation which becomes very visible once the annotation of mentions includes the semantic class is the treatment of metonymy, as in the following example:

(3.5) *Paris* rejected the "logic of ultimatums".

The meaning of the example could be interpreted roughly as

(3.5') A French government official made a statement to the effect of a French official position of disapproval regarding the "logic of ultimatums".

The problem for any coreference guideline is not to specify whether the mention *Paris* in (3.5) would be coreferent to other mentions of:

1. The city of Paris
2. The country of France (as a geographic extension)
3. The French government *or*

[6] These corpora are available from the Linguistic Data Consortium as:
Alexis Mitchell et al. (2003): ACE-2 Version 1.0, LDC2003T11
Alexis Mitchell et al. (2005): ACE 2004 Multilingual Training Corpus, LDC2005T09
Christopher Walker et al. (2006): ACE 2005 Multilingual Training Corpus, LDC2006T06

4. The government official uttering the sentence

Different annotation guidelines offer different (partial) solutions for this problem: The ACE guidelines resolve the ambiguity between 2 and 3 by assuming *geopolitical entities* (GPEs), i.e., a conflation of a country, its government, and its inhabitants in an effort to mitigate the problems created by polysemy.

In contrast to ACE's method of dealing with this specific problem type, the more recent OntoNotes corpus (Pradhan *et al.*, 2007) incorporates guidelines based on the diametrally opposite solution: It does not treat the government, the area, and the population of a country as different aspects of the same geopolitical entity; instead, it prescribes that metonymous uses are distinguished from non-metonymous uses for the purpose of coreference annotation; thus, in a document that contains the sentences

(3.6) [$_1$ South Korea] is a country in southeastern Asia. ... [$_2$ South Korea] has signed the agreement.

it is necessary to distinguish between "South Korea" mentioned as a country and "South Korea" as a metonymous mention of the South Korean government.

Both solutions make it necessary that the preprocessing assigns semantic classes to all mentions, and that these semantic classes are consistent with the specification of the coreference corpus (including metonymy resolution in the case of OntoNotes).

3.1.4 Coreference annotation in the TüBa-D/Z corpus

For the research reported in this thesis, our interest falls mainly on the syntactically and referentially annotated TüBa-D/Z corpus (Hinrichs *et al.*, 2005a).

The common existence of syntactic and referential annotation has many advantages with respect to corpora such as the MUC and ACE corpora, which are not backed by a treebank. In the annotation of the TüBa-D/Z, markables and their boundaries are extracted automatically and do not have to be marked manually by the annotator; annotation of minimal spans, which would otherwise be necessary to ensure that system-generated markables can be aligned to the gold standard, is not necessary since minimal spans can be automatically derived from treebank trees.

The annotation is carried out according to an annotation scheme inspired by the MATE proposal for coreference annotation. (Identity) coreference relations, which would receive the "*ident*" relation label in the MATE are subdivided into three categories (*anaphoric*, *cataphoric*, and *coreferent*) based on the following criteria:

- *anaphoric* links a pronominal anaphor (personal, demonstrative, reflexive, and relative pronouns, including possessive pronouns) to its closest (preceding-in-linear-order) antecedent.

- *cataphoric* links a pronoun to the mention it is resolved to if that mention comes later in the linear order of the text.
- *coreferent* links subsequent-mention definite noun phrases to the closest preceding mention. The label is also used for definite description that occur in copula clauses and similar constructions (for which the MATE guidelines use the term *predicate nominals* and allow their annotation if the copula clause expresses explicit equality, whereas van Deemter and Kibble reject the notion that such cases are coreference relations).

  An example of such a definite description occurring as a predicate nominal can be seen in the following example:[7]

  (3.7) [$_2$ Er] ist [$_2$ nicht nur das musikalische, sondern auch das romantische Rückgrat des Trios].
  [$_2$ *He*] *is* [$_2$ *not just the musical, but also the romantic backbone of the trio*].

  As discussed in subsection 2.2, the definite article in such noun phrases does not signal a subsequent mention of an introduced discourse referent (as these noun phrases do not refer), but depends on semantic uniqueness properties of the description itself (as in the case of superlatives, or functional concepts).

For the distinction between *anaphoric* and *cataphoric*, syntactic relations overrule the linear order of the text in cases of first-person pronouns and direct speech (in which case the pronoun has the NP in the speaker role as its anchor), as in the following example:[8]

(3.8) "Sollen sie [$_1$ mich] als Narren ansehen!" [$_1$ Ulrich Görlitz] diente freiwillig in Hitlers Armee.
"Let them consider [$_1$ *me*] *a fool!"* [$_1$ *Ulrich Görlitz*] *served in Hitler's army voluntarily.*

A third relation, *bound*, that is used for anaphora bound by the same quantifier as their antecedent:

(3.9) ... [$_1$ jeder] auf [$_1$ seine] Weise ...
... [$_1$ *each*] *in* [$_1$ *his*] *own way* ...

The TüBa-D/Z annotation also includes two non-identity relations that are inspired by those in MATE's extended scheme:

- *split_antecedent* links a plural pronoun to the descriptions that together refer to the summum that the pronoun refers to, as in cases such as the following example:

[7] TüBa-D/Z, sentence 2181.
[8] TüBa-D/Z, sentences 2103ff.

(3.10) [$_1$ John] met [$_2$ Mary]. Then [$_{3\rightarrow1+2}$ they] went to the cinema.

In this case, *they* refers to the summum of *John* and *Mary* and a *split_antecedent* relation is established from *they* with both *John* and *Mary* as targets.

- *instance* relations are annotated between a set-denoting noun phrase and a first-mention noun phrase that denotes a member of the set:[9]

(3.11) Auch [$_1$ prinzipielle Fragen], wie [$_{2\rightarrow1}$ die nach der Mandatierung künftiger Nato-Einsätze durch die UNO], werden in dieser grundsätzlichen Form nur in Deutschland diskutiert.
Further, [$_1$ *questions of principle*], *such as* [$_{2\rightarrow1}$ *the one of future Nato interventions being mandated by the UNO*] *are only ever discussed in this fundamental form within Germany.*

The TüBa-D/Z also includes explicit marking of expletive *es* pronouns of multiple kinds: So-called "Vorfeld-*es*", which do not have a grammatical function, receive the ES grammatical function label in the syntactic annotation, whereas others (both correlates of clausal arguments and semantically empty subjects of weather verbs) receive the *expletive* label.

Reflexive pronouns are differentiated between regular reflexive pronouns, which receive a thematic role from the verb they occur with, and so-called *inherent reflexives*, which are a semantically empty but obligatory verb argument for some verbs such as *sich schämen* (to be ashamed).

A study of inter-annotator agreement for the coreference relations (*anaphoric*, *cataphoric* and *coreferent*) can be found in (Versley, 2008b). Generally, the inter-annotator agreement for TüBa-D/Z is quite good, at $F = 0.85$, or $\kappa = 0.81$. This compares favorably to other current efforts (Poesio and Artstein, 2008 report α values of $0.6 - 0.8$ for agreement between different annotators in the ARRAU corpus; Hendrickx *et al.*, 2008 report $F = 0.78$ for single anaphor-antecedent links).

3.1.5 Other corpora

Besides the English-language corpora mentioned above – the MUC and ACE corpora, and the OntoNotes corpus, large coreferentially annotated corpora exist for a large number of languages (see table 3.1).

Many of these corpora are inspired by the MUC or MATE guidelines: the Dutch COREA corpus (Hendrickx *et al.*, 2008) is loosely based on the MUC guidelines, but offers separate relation types for predicative relations, and also includes bridging relations. The ARRAU corpus (Poesio and Artstein, 2008) closely follows the MATE guidelines, including some of the relations from the 'extended scheme'. The Spanish/Catalan AnCora corpus (Recasens *et al.*, 2007) is based on the MATE

[9] TüBa-D/Z, sentence 19865

Name / Scope	Language	Reference	Size (words)
ACE-2005	Arabic	Walker *et al.* (2006)	100k
OntoNotes		Weischedel *et al.* (2008)	100k
ACE-2005	Chinese	Walker *et al.* (2006)	≈200k
OntoNotes		Weischedel *et al.* (2008)	550k
COREA	Dutch	Hendrickx *et al.* (2008)	325k
MUC-6	English	Grishman and Sundheim (1995)	30k
MUC-7		Chinchor (1998)	30k
ACE-2005		Walker *et al.* (2006)	400k
OntoNotes		Weischedel *et al.* (2008)	500k
Arrau		Poesio and Artstein (2008)	90k
DEDE (definite descriptions)	French	Gardent and Manuélian (2005)	50k
Potsdam Commentary Corpus	German	Stede (2004)	33k
TüBa-D/Z		Hinrichs *et al.* (2005a)	630k
Venex	Italian	Poesio *et al.* (2004b)	40k
i-Cab		Magnini *et al.* (2006)	250k
NAIST Text Corpus	Japanese	Iida *et al.* (2007)	38k sentences
AnCora-ES	Spanish	Recasens *et al.* (2007)	400k
Tusnelda (B11)	Tibetan	Wagner and Zeisler (2004)	<15k

Table 3.1: Referentially annotated corpora in different languages

scheme but adds additional relation types for lists and for larger situation (*contextual*) descriptions which have to be interpreted with respect to the spatial/temporal coordinates of the article (e.g. descriptions such as *this year*). Very few corpora take the ACE guidelines as inspiration for coreference annotation; the Italian iCab corpus (Magnini *et al.*, 2006), where ACE-style annotation of semantic classes is augmented with (ACE-inspired) coreference annotation, is a notable example.

While the MATE proposal for the annotation of empty subjects and clitics is adequate for covering these phenomena in the Romance languages that have them (e.g., Italian and Spanish), other languages such as Chinese, Japanese and Korean can have zero pronouns in object positions. As a result, corpora such as the Japanese NAIST Text Corpus (Iida *et al.*, 2007) use an annotation scheme that is more adapted to the frequent occurrence of zero arguments (and modifiers) to verbs and nominalizations.

3.2 Evaluation

The dominant use of gold standards, besides inductive learning in machine learning systems, is the use of a **quantitative evaluation measure** to track improvements to a system, to compare different systems, or as a yardstick to judge the fitness for a given application.

Empirical evaluation, with a formalized setting and standardized measures, is

often seen as an important part in the establishment of a task, and subsequent research is expected to report evaluation measures and, possibly, to quantify improvements in terms of evaluation figures in that setting.

It has to be said that using only quantitative evaluation results carries a danger of numerical improvements that are not reflected in application usefulness, which means that evaluation-driven development should be accompanied by a critical discussion of the relation between **evaluation setting** (the input, and specification of the desired output, of a system) and **evaluation metric** (the method used to summarize the matching between system output and gold-standard annotation) on one hand, and potential usefulness to an application on the other side.

Different evaluation methods (i.e., evaluation settings and evaluation metric) may also be differently forgiving (or nearly insensitive) to certain kinds of errors, which means that, depending on the evaluation measure, it is not always safe to assume that a higher evaluation score results in better results in the application setting (cf. the results of Callison-Burch *et al.*, 2006 for machine translation or Miyao *et al.*, 2008 for parsing).

To fully understand reported results in coreference resolution, it is therefore helpful to understand both the influence of different evaluation settings, and also the behavior of different evaluation metrics when it comes to different system responses.

Link-based and set-based measures (discussed in subsections 3.2.1 and 3.2.2) count true positives (correctly established coreference decisions), false positives (incorrectly established coreference decisions) and false negatives (wanted coreference decisions that are absent from the system response).

In both kinds of measures, **Precision** is the ratio of true positives to the sum of true positives and false positives (in a link-based measure, the total number of coreference links created by the system); **Recall** is the ratio of true positives to the sum of true positives and false negatives (in a link-based measure, the total number of coreference links contained in the gold annotation).

$$\text{Precision} = \frac{\#\text{correct}}{\#\text{resolved}} \qquad \text{Recall} = \frac{\#\text{correct}}{\#\text{wanted}}$$

To summarize precision and recall into one evaluation measure, it is common to use the harmonic mean of precision and recall, called *F-measure* following its introduction as evaluation measure for information retrieval by van Rijsbergen (1979). The harmonic mean of two numbers is well approximated by the arithmetic mean of these two numbers when they are close to each other; when the difference is large, the harmonic mean is closer to the minimum of the two numbers. The F-measure is closely related to other measures that focus on the behavior regarding one minority class, such as the *positive specific agreement* measure used by Spitzer and Fleiss (1974).

$$\text{F} = \frac{2}{1/\text{Precision} + 1/\text{Recall}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 \times \#\text{correct}}{\#\text{resolved} + \#\text{wanted}}$$

However, link-based and set-based systems calculate the numbers of true positives, false positives, and false negatives differently.

- A *Link-based measure* scores single resolution decisions, and not resolving a discourse-old mention at all (one false negative), would be punished less than resolving it to an incorrect antecedent (one false negative, and one false positive).

- The MUC measure of Vilain *et al.* (1995), which is a *set-based measure*, compares the partition imposed by the coreference chains, and counts (roughly speaking) the number of splitting or merging operations necessary to turn the system response into the gold standard.

  In the case of the MUC measure, the numbers #wanted and #resolved still correspond to the count of coreference links in the gold standard (wanted) and in the system response (resolved). However, the numerator (correct) is calculated more optimistically than in a link-based measure: A wrong non-coreferent link that the system creates is counted as a true positive whenever the two coreference chains from the gold standard have already been erroneously linked. Accordingly, the number of true positives in the MUC measure is the maximum number of correct links that can be obtained with *any* assignment of links that corresponds to the coreference chains (see the explanation in subsection 3.2.2 and figure 3.2, *below*).

- Finally, *alignment-based metrics* such as the ECM/CEAF score of Luo (2005) force a one-to-one alignment between system-generated and gold standard coreference chains, yielding no reward for partial chains of correctly identified links if they are not joined with the majority chain.

  Alignment-based metrics do not offer separate measures for precision and recall, but implicitly punish any system that creates vastly less or more coreference chains, since a portion of the coreference chains will then remain unaligned.

3.2.1 Link-based measures

A link-based measure is the simplest imaginable measure for success in coreference: a coreference link introduced by the system is judged as correct whenever it links two mentions that are coreferent according to the gold standard.

It is important to note that the system does not need to reproduce the exact set of links that are annotated in the gold standard: among a set of M coreferent mention, any cycle-free set of $M - 1$ links (i.e., any spanning tree) among these mentions will yield the same coreference set. In the common case where every mention is resolved to at most one previous mention, no cycle can occur.

Early versions of the MUC coreference task definition (up to version 2.0) simply calculated precision and recall of the links in the system response with respect

to the links in the annotated reference. Requiring the system to produce the exact links (and scoring links to other antecedents which are correct but not the closest as wrong) quickly leads to problematic cases. Consider an example such as

(3.12) [A1 Peter] likes sweets.
[A2 The boy] has always liked chocolate,
but at the moment, [A3 he] is devouring it like mad.

The gold annotation includes the links $\langle A_1\text{–}A_2,\ A_2\text{–}A_3\rangle$. If a system simply linked *he*(A_3) to *Peter*(A_1), ignoring *The boy*(A_2), the resulting system response $\langle A_1\text{–}A_3\rangle$ would have been scored as completely wrong, with no partial credit given (cf. figure 3.3).

Evaluating links as correct if the antecedent is coreferent with the anaphor solves this problem since the link $\langle A_1\text{–}A_3\rangle$ would be counted as correct. The provision of cycle-freeness would be needed to exclude system responses such as $\langle A_1\text{–}A_2,\ A_2\text{–}A_3,\ A_1\text{–}A_3\rangle$, which contains three correct links, but form a cycle (whereas a non-cyclic structure can only yield two correct links, which matches the number of links in the gold standard).

Link-based measures can easily be used to get a differentiated picture of the resolution performance of a system for different kinds of expression (e.g., pronoun resolution versus names versus nominals, or third-person versus other pronouns).

3.2.2 Set-based measures

In the development of the MUC coreference task, the initial proposal for exact matching of links was discarded for a different kind of metric, which we will call a *set-based measure* here. A link-based metric, while it does not yield the undesirable behavior of exact link matching in the example in figure 3.3, would still assign different scores to link structures that result in the same equivalence classes (cf. figure 3.2). The following (somewhat contrived) example shows this:

(3.13) [A1 Peter] burned the pancakes out of distraction.
[B1 The milkman] came by and
[A2 he] thought about buying some milk.

Resolving both B_1 (*The milkman*) and A_2 (*he*) to A_1 (*Peter*) would then get a better score than resolving B_1 (*the milkman*) to A_1 (*Peter*) and A_2 (*he*) to B_1 (*the milkman*), even though they both result in the same partition $\{\{A_1, A_2, B_1\}\}$.

Vilain *et al.* (1995) propose a solution to both the former and the latter problem by introducing precision and recall statistics for equivalence classes (see the "Sets" lines in the examples 3.2 and 3.3), which was adopted for the MUC coreference task definition 2.1 and later.

To determine precision and recall in a set-based fashion, the MUC measure uses a 'number of links' statistic for a given partitioning, and a method to compute the intersection of two partitions:

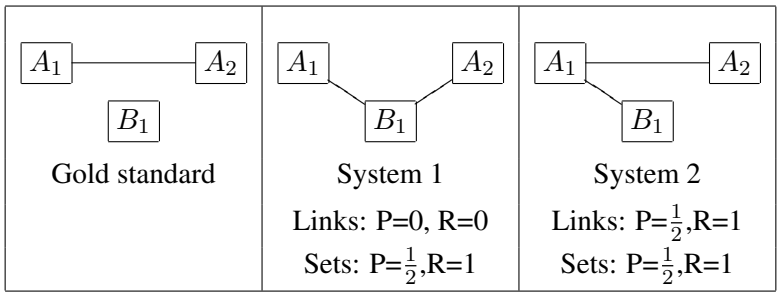


Figure 3.2: Link-based evaluation: Example 1

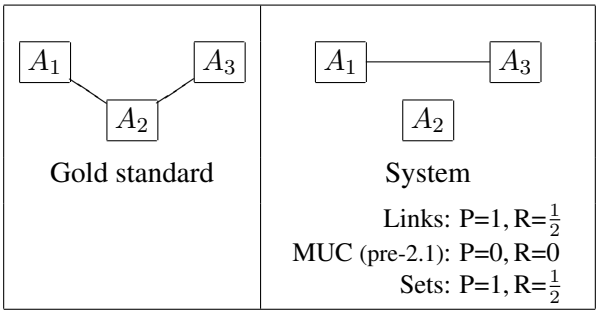


Figure 3.3: Link-based evaluation: Example 2 (from Vilain *et al.*)
MUC (pre-2.1): link-based score from MUC coreference task definition up to version 2.0

If we write partitions as sets of equivalence classes, for example as $G = \{\{A_1, A_2\}, \{B_1\}\}$ for the gold standard sets of figure 3.2, we get the number of equivalence classes

$$|G| = 2$$

and the number of class members

$$\left|\bigcup_{A \in G} A\right| = \left|\{A_1, A_2, B_1\}\right| = 3$$

For any partition G, we may define $\ell(G)$ as the '**number of links**', or equivalently the number of elements in a set minus the number of equivalence classes in S:

$$\ell(G) := \left|\bigcup_{A \in G} A\right| - |S| = \sum_{A \in G} |A| - 1$$

(The 'number of links' would be $3 - 2 = 1$ in our example, corresponding to our intuition that we have only one link which connects A_1 and A_2).

We can also define, for partitions S_1 and S_2, a partition $S_1 \sqcap S_2$ which contains the **intersection of equivalence sets** to S_1 and S_2:

$$S_1 \sqcap S_2 := \{A_1 \cap A_2 \,|\, A_1 \in S_1 \wedge A_2 \in S_2 \wedge A_1 \cap A_2 \neq \varnothing\}$$

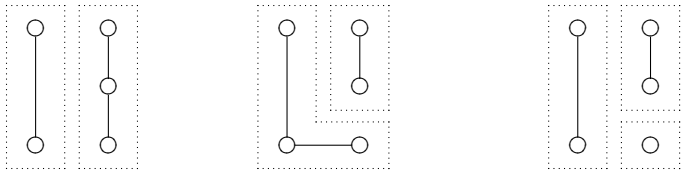


Figure 3.4: Examples for the ℓ (link count) function on partitions

Then, for a mention partition G of the system annotations and a mention partition S of the system response, we can define precision and recall in a straightforward way:

$$P = \frac{\ell(G \cap S)}{\ell(S)} \qquad R = \frac{\ell(G \cap S)}{\ell(G)}$$

In the first example, we would have $G = \{\{A_1, A_2\}, \{B_1\}\}$ and $S = \{\{A_1, A_2, B_1\}\}$. This would give us $\ell(G) = 1$ and $\ell(S) = 2$, and, with $G \cap S = \{\{A_1, A_2\}, \{B_1\}\}$, $\ell(G \cap S) = 1$, $P = \frac{1}{2}$ and $R = 1$.

Note that the addition of singleton mentions to a partition does not change its link count, and, accordingly, the addition of singleton mentions to gold standard and/or system response would not change the precision or recall scores.

Other scoring algorithms readily exist: one of the better known, Bcubed (Bagga and Baldwin, 1998a), includes singletons in the evaluation. A clustering measure used in general clustering tasks, purity, could be used together with its dual (inverse purity) to score partitions as well.

It is possible to cast all three (MUC, Bcubed, purity) in a common framework, starting with the notion that recall in MUC and Bcubed is just the dual of precision: given a precision measure that scores how much of the equivalence classes in a system output is part of the 'consensus' between system and gold standard, one can also ask how much of the information in the gold standard is part of the consensus, yielding a recall measure.

For some measure μ_1 that monotonically increases with cluster size and fulfills $\mu_1(A \uplus B) \geq \mu_1(A) + \mu_1(B)$, we get a **score for one cluster**:

$$score_0(A, G) = F\left(\left\langle \frac{\mu_1(A \cap B_i)}{\mu_1(A)} \middle| B_i \in G \right\rangle\right)$$

where F is a function that maps a vector of non-negative values with a sum less or equal to one to a single value in the interval $[0, 1]$ (which should obviously map unit vectors $\langle 0, \cdots, 1, \cdots, 0 \rangle$ to 1 and others to a smaller value).

The **score for the partition** can then be calculated as a weighted sum of the scores for all clusters:

$$score(S, G) = \frac{\sum_{A \in S} \mu_2(A) \cdot score(A, G)}{\sum_{A \in S} \mu_2(A)}$$

For the MUC score, we can set $\mu_1(A) = \mu_2(A) = \max(0, |A| - 1)$ and $F(x) = \sum_i x_i$. We can also derive the Bcubed score if we set $\mu_1(A) = \mu_2(A) = |A|$, with $F(x) = \sum_i x_i^2$. A variant of Bcubed, called Bcubed-IR by Bagga and Baldwin, uses a constant cluster weighting of $\mu_2(A) = 1$. Cluster purity can be calculated with $\mu_1(A) = \mu_2(A) = |A|$ and $F(x) = \max_i x_i$.

Seen in this common framework, several properties of the MUC, Bcubed, and cluster purity measures become apparent:

- With respect to one case where a single element is split off a cluster and a cluster is split in two halves, the MUC measure, just like link-based measures, treats the two cases equally, whereas Bcubed and cluster purity punish the latter case more harshly.

- If we double the number of mentions by simply cloning each mention within a cluster (i.e., instead of just one element which is part of a cluster, we would have two), the MUC score for a given system response will increase (since the number of correct links has increased). The Bcubed and cluster purity measures will not increase as the relative sizes of clusters will be unaffected.

- Cluster purity is not affected by differences in the smaller (i.e., non-majority) components in a cluster: for a system response that merges all mentions, it will assign the same score regardless whether the gold partition was $\{\{A_1, A_2\}, \{A_3, A_4, A_5\}\}$ or $\{\{A_1\}, \{A_2\}, \{A_3, A_4, A_5\}\}$, since the majority class of A_3 – A_5 determines the maximum. Bcubed and the MUC score will give a lower value to the second case.

- Bcubed, cluster purity and other measures based on set size (as opposed to link count), including Luo's ECM/CEAF measure (which is detailed in the next subsection), encounter a problem when systems create mentions that are not part of the gold standard: if these mentions are added to the gold standard as singletons, the responses of systems that create a different number of singleton mentions are not comparable to each other (as the newly created singletons will artificially inflate the agreement score).

This common framework can also be used to create a measure (named 'link purity' in table 3.5) that, like the MUC measure, is based on 'link counts' (i.e., insensitive to the addition of singletons), yet, like purity, is more sensitive to over-merging: combining link-count as cluster heaviness score $\mu_1(A) = \mu_2(A) = |A| - 1$ and using the maximum (instead of the sum, as in the MUC score), for scoring: $F(x) = \max_i x_i$. As can be seen in 3.5, this measure follows the MUC measure closely in most cases, but punishes over-merging more harshly (but only in terms of precision, as opposed to the alignment-based measures presented in the next subsection).

		MUC			Link Purity			ECM/CEAF		
Gold	$[A_1\ A_2\ A_3]\ [A_4\ A_5]$	Prec	Recl	F	Prec	Recl	F	Prec	Recl	F
System 1	$[A_1\ A_2\ A_3\ A_4\ A_5]$	3/4	3/3	0.86	2/4	3/3	0.67	3/5	3/5	0.60
System 2	$[A_1\ A_2]\ [A_3\ A_4\ A_5]$	2/3	2/3	0.67	2/3	2/3	0.67	4/5	4/5	0.80
System 3	$[A_1\ A_2]\ [A_3]\ [A_4\ A_5]$	2/2	2/3	0.80	2/2	2/3	0.80	4/5	4/5	0.80
System 4	$[A_1]\ [A_2\ A_3\ A_4\ A_5]$	2/3	2/3	0.67	1/3	2/3	0.44	3/5	3/5	0.60

Figure 3.5: Behavior of MUC, Link Purity and ECM/CEAF on different examples

3.2.3 Alignment-based measures

The scoring method of Vilain *et al.* is designed to be an optimistic generalization of link-based measures to coreference sets, with the MUC scores being the best attainable scores for any decomposition into links of system and gold standard partitions. While this design leads to some non-intuitive effects on the small scale (misclassifying one mention into the wrong coreference set counts as one precision and one recall error, while completely merging two coreference sets counts as a single recall error), these effects are compounded when evaluating the system response on true (gold) mentions, where all singletons and non-referring mentions are removed: In this case, just merging *all* mentions into a single coreference chain simply incurs a number of precision errors corresponding to the number of (gold-standard) coreference chains (minus one), whereas the number of correct links is evaluated as the total number of gold mentions (minus one), thus giving a trivial baseline with 100% recall and usually about 80% precision.

While this counter-intuitive result is not solely due to the MUC evaluation metric (see the discussion at the end of this section), several researchers (Trouilleux *et al.*, 2000; Luo, 2005) propose solutions that handle these cases more gracefully by punishing over-merging more strongly. Especially Luo's proposal, variously called ECM(-F) or CEAF, is widely used in practice since it avoids some drawbacks of the MUC measure and corresponds to the intuitive notion that the system response should (faithfully) reproduce the sets of mentions corresponding to each discourse referent.

Common to both measures is the use of weighted links between system and gold coreference sets, that are then used to induce an **alignment** by selecting those links such that (i) every coreference chain to the system corresponds to at most one coreference chain from the gold standard, and vice versa, and (ii) the highest weight among these assignments is reached.

Using a one-to-one alignment, overmerging (or any case where two coreference chains are completely merged, rather than having single elements misclassified) is punished more harshly in terms of precision (since the merged entity will only receive credit for covering one gold-standard coreference chain, and not the others),

but also in terms of recall (since the merged entity will only be aligned to one gold-standard coreference chain and recall for the others will be zero) – to illustrate this, see figure 3.5, which compares Vilain et al's MUC measure, link purity, and the ECM/CEAF measure.

Trouilleux *et al.* compute the weights of the links as the weighted proportion between the sets mentions that are in the intersection and in the union of system and gold coreference chain (Dice coefficient):

$$\text{score}(A_i, B_j) = \frac{2|A_i \cap B_j|}{|A_i| + |B_j|}$$

This score has the property that the sum of all possible links for one coreference chain is always at most one, also in the case where the score weights names, common noun phrases and pronouns differently. (Trouilleux *et al.* propose a weighting of 0.6 for names, 0.3 for common noun phrases, and 0.1 for pronouns). They then compare the number of "correct" links that are common to the aligned coreference chains with (i) the link count for the system's coreference chain, to get the precision, and (ii) the link count for the coreference chains in the gold standard, to obtain the number for recall.

In a similar fashion, the "entity-constrained" metric proposed by Luo *et al.* (2004); Luo (2005), variously abbreviated ECM-F (for Entity-Constrained Metric) or CEAF (Constrained Entity-Aligned F-Measure), first computes an alignment and then compares the sets of mentions in the system's (for precision) or the gold standard coreference chains to the result of the alignment; to be counted as correct, a mention must be in two (system and gold standard) coreference chains linked by the alignment.

Another mention, which predates Luo's CEAF proposal (and could be said to have motivated it) is the scoring used in the Automatic Content Extraction (ACE) evaluations. Like CEAF and Trouilleux *et al.*'s measure, the ACE measure creates a one-to-one alignment between coreference chains ("entities") in the gold annotation and the system response. Like Trouilleux *et al.*'s measure, and unlike CEAF, the ACE measure assigns different weights to different kinds of mentions, with name mentions being weighted the most heavily and pronoun mentions receiving a fairly low weight.

The details of the ACE evaluation measure cannot be described in full here since it uses a multitude of weighting factors and the exact definition has been changed between successive evaluation to better fit the goal of being an evaluation measure that is predictive of performance in information extraction. As it uses weighting both for computing an alignment and for the computation of the score (unlike Trouilleux et al., who only use it for computing the alignment), Luo criticizes that the ACE score tends to over-emphasize name matching, while progress on pronoun resolution is poorly reflected in the score: its weighting puts an emphasis on named entities and de-emphasizes pronouns which, means that just matching names assures a result above 86% when evaluated on gold mentions.

3.2.4 Properties of Evaluation Metrics

Evaluation metrics are differently sensitive to various kinds of errors, and therefore the evaluation metric also has a marked influence on the reported usefulness of techniques, as under- or over-merging (producing larger or smaller coreference chain than in the gold standard) are differently handled by these metrics.

As an example, consider the coreference chains in the following example (3.14):

(3.14) a. The white horse ... the horse
b. The red car ... the car
c. The green car ... the car

A system that is unsure about the anaphoric mentions with "*the car*" - let us further assume that other heuristics (such as recency) also do not help here and the result would come down to guessing - can either abstain from the resolution of the anaphoric definite description, leaving the mentions alone instead of inserting them in an existing coreference chain (*abstention/under-merging*). It could also insert them into one of the coreference chains, even if it is not sure about it (*random guessing*). Or it could simply merge both coreference chains including "*the red car*", "*the green car*" and both "*the car*" mentions in a single chain (*merge all*).

A **link-based measure** would count an incorrect coreference link (one false negative and one false positive) more than not resolving a discourse-old mention at all (one false negative only). Therefore, depending on the expected number of correct decisions, it is plausible that link-based measures would encourage **abstentions** or random guessing more than over-merging.

In the case of the **MUC measure**, the number of true positives is calculated more optimistically than in the link-based case (recall the example of figure 3.2 on page 68).

As a result, **over-merging of mentions that belong to some kind of equivalence class** (e.g., all "*he*" pronouns) is preferred by the MUC metric over random resolution. In the case of completely merging N coreference chains (containing M mentions altogether, with M significantly greater than N), only $N - 1$ false positives are counted with the MUC metric, whereas not linking them at all yields $M - 1$ false negatives, and random linking would result in some combination of those false negatives and false positives.

The **ECM/CEAF metric** punishes both under-merging and over-merging, which means that the preferred strategy to reach a good ECM/CEAF score would be to produce **roughly the same number of coreference chains** in order to avoid a larger number of non-aligned coreference chains.

The behavior of these metrics should be considered with the requirements of an application in mind: Hendrickx *et al.* (2008) note that a high-recall coreference system that leads to an increase of 50% in the relations found by an information extraction system only yields minimal improvement in F-measure due to the low

precision of the high-recall resolver.

In a similar vein, Stuckardt (2003) argues that "*Precision errors are critical, since they can lead to a wrong answer derived from a context of a non-coreferring occurrence, including an incorrect substitute expression. Thus, precision errors can be expected to cause more potential damage with respect to the applications TS and QA than recall errors.*"

If one wants to argue not from the perspective of the evaluation as a more well-behaved proxy for application usefulness, but rather from a system development perspective, it would be preferable to use an evaluation metric that closely corresponds to single decisions of the system (in the most common case of a resolver that finds a plausible antecedent for a potentially-anaphoric mention, resolution decisions), which again could indicate a preference for link-based over other evaluation metrics.

Independently of these concerns, the metrics are in general not usable in every case: If the system creates coreference clusters directly without constructing antecedent chains, link-based measures are unapplicable. If the number of markables can vary between the gold standard and the system response (as a result of preprocessing, if it is not based on a fixed point of reference such as noun phrases in the treebank), alignment-based measures such as ECM/CEAF are not applicable.[10]

3.2.5 Evaluation setting

A common question with both algorithms and systems for computational anaphora resolution is how good these perform or if one is better than the other, preferably answered in hard, concrete numbers. Many of the earlier approaches for pronoun resolution give numbers of the type "*N*% of the pronouns are resolved correctly", which seems exactly like the answer one wants, until thrown onto the fact that *what* is being compared is not always the same. The discussion of different evaluation metrics for coreference resolution has for a long time obscured the equally important question of the evaluation setting, which has however been adequately discussed for the problem of evaluating pronoun resolution. In the latter context, Mitkov (2000) and Byron (2001) point out the following central questions:

- Is this an evaluation of the algorithm (i.e., assuming perfect preprocessing, including agreement features like number or gender) or of a whole system, where preprocessing steps such as parsing and determination of gender features is done automatically?

- Does the evaluation include or exclude difficult cases such as first-person pronouns (which may not be resolvable to an antecedent), cataphora, cases

[10] Researchers using Bcubed and ECM/CEAF with system markables have created variants of the measures that take this into account. Unfortunately, the solutions chosen by different researchers rarely allow a meaningful comparison between published results using different variants of Bcubed or ECM/CEAF.

of expletive pronouns, or pronouns and demonstratives that refer to clauses instead of noun phrases?

- On which type of texts is the evaluation carried out, as technical manuals seem to be easier to treat with pronoun resolution than newspaper text?

Quantitative evaluation results in general – be it for pronoun resolution or for coreference resolution at large – have been shown to vary according to a variety of factors:

- In a given setting, different *evaluation measures* may be differently tolerant towards over- or under-merging. Luo (2005) found that in a setting he investigated – gold mentions on the MUC 6 corpus, the MUC measure had a strong bias towards over-merging.

- Depending on the annotation guidelines, coreference resolution on different corpora can mean that the task to be solved consists in different phenomena. For example, the MUC-6/MUC-7 corpora see the detection of apposition relations as part of the coreference resolution task, whereas the TüBa-D/Z considers the complete noun phrase as one mention (including both parts of the apposition) since the information on appositions is already contained in the syntactic information.

- System performance also strongly varied depending on the actual text or text genre that is used as a basis. Genre-specific variations on the distribution of different kinds of anaphora, as well as factors such as text length means that research using the ACE corpus usually shows score variation across the broadcast news, newswire, and newspaper sections. For German, Kouchnir (2004) finds a large difference for the pronoun resolution results between the simpler prose of the Heidelberg Text Corpus and Spiegel newspaper text.

As a result, evaluation figures obtained on different annotated corpora, or using different evaluation procedures, cannot easily be compared. These concerns (scope of the task and the text genre used) seem to be less of an issue when looking at results from using well-known corpora that have been used for common evaluation. However, some care is still necessary in the interpretation of numeric results when different evaluation metrics and evaluation settings are used.

Most research subscribes to the end goal of achieving both a system (i.e., a usable, working implementation) *and* an algorithm (i.e., a general idea of how to solve the task). Considering this goal, a system evaluation would be at least as important as an isolated evaluation of an algorithm. However, it is often sensible to start with a simplified task to work around limitations of the used preprocessing techniques, or scalability problems in some part of the approach, while still reaping the benefits of quantitative evaluation and data-driven development. Most often, such simplifications consist in assuming a certain kind of information as given,

or choosing to use only a subset of the data that is more well-natured in certain respects.

In a coreference system, much of the complexity in evaluating a system comes in the form of letting the system create its markables and having to reconcile the differences between system markables and annotated (gold) markables in the evaluation.

Requiring the system to create its own markables thus results in two complications: One is that the **markable extraction** itself can be error-prone in the case of a complicated markable definition as in MUC-7 (where certain premodifiers can be markables, but chains that consist only of premodifiers are to be discarded), but also due to imperfections of the preprocessing components. Markable extraction will then be evaluated in conjunction with the rest of the system, resulting in a score that results both from the (ostensibly non-interesting) markable extraction and coreference proper.

The other drawback is that in the case of **non-exact matches**, as in {*President [Bill Clinton]*} (where the {} denote the gold markable and the [] denote the markable extracted by the system), some heuristic has to be used to match system markables, further complicating the process of evaluation. The only feasible solution for non-exact matches that has been found is the one taken in the MUC and ACE evaluation, where the gold-standard markables are marked both with a *minimal* (i.e., head or chunk) and a *maximal span* (i.e., including all modifiers); system markables are matched to a gold-standard markable if their span includes the minimal span of that markable, but is included by its maximal span.

As a simpler alternative, it is possible to use the markables from the corpus which are part in coreference relations, or a *well-defined superset* of the gold-standard markables, such as the set of all (maximal) noun phrases in the case of a treebank.

In the case of a referentially annotated corpus backed by a treebank, or one that generally identifies every noun phrase in the corpus, such a simplification is straightforward and defensible especially since it eliminates the necessity of aligning system and gold-standard markables and allows the use of alignment-based metrics.

It is usually quite defensible to use the same methods of quantitative evaluation also for the simplified settings, as long as the fact is made transparent. However, it has to be believable that a system actually could work in that way; and settings that demonstrably yield misleading evaluation results must be avoided.

Using only markables that were annotated in the gold standard (in the setting called *No Singletons* below), may, and usually does, overestimate the possible accuracy for eliminating markables that do not take part in coreference relations, such as expletives, other non-referring noun phrases, or simply entities that are only mentioned once. Such a setting yields near-perfect information on whether a mention is coreferent with any other mention or not, since these singleton mentions (which constitute about 50-60% of all NPs) are already removed from considera-

tion.

Among others, the literature on coreference resolution uses the following simplified settings to derive evaluation results:

- *No Singletons*: The system only gets to see markables that are part of a coreference chain in the gold standard. As the coreference annotation (almost always) only marks the span of mentions that are part of a coreference chain for practical reasons, this way of evaluating leaves out more than half of all possible antecedents (i.e., it makes the choice of a correct antecedent considerably easier), as well as not proposing singletons as potentially anaphoric mentions to be resolved (which makes the choice between resolving or not resolving a potentially anaphoric description much easier).

- *Perfect Anaphoricity Information*: The system only gets to resolve those markables that are actually discourse-old; some researchers (e.g., Garera and Yarowsky, 2006) only include markables that are discourse-old and have an antecedent in a small window of few preceding sentences.

- *Gold-Standard Subset*: The system is used on markables as identified in the gold-standard, but the gold standard also contains singletons. In the case of the ACE evaluation and the corresponding corpora, for example, all noun phrases belonging to certain semantic classes are marked.

- *Disregard Spurious Links*: Some researchers (Ji *et al.*, 2005) use automatic markable information, but remove all non-matched markables from the system response and also remove all non-identified gold markables from the answer key. The net effect of this is a slightly greater distortion than in the *No Singletons*/*Gold-Standard Subset* settings, since not only singleton markables are left out of the evaluation, but also irregular markables that would be hard to identify (and possibly also difficult to use in coreference resolution).

The net effect of such simplified task settings – we will refer to the *No Singletons*, *Gold-Standard Subset*, and *Disregard Spurious Links* settings as "*true mentions*", as this expression is usually used in the literature, is that the coreference system will need to eliminate substantially fewer spurious antecedent candidates.

The overview in chapter 4, as well as the experiments in chapters 6 and 7 will further expound the notion that different evaluation settings results not only in large differences in score, but also in the demonstrable usefulness of high-recall, low-precision information sources for the coreference resolution of definite noun phrases.

Evaluation on gold-standard mentions has the problem that non-links between mentions are no longer in the minority, which emphasizes the weakness of the MUC score in not being sensitive enough against over-merging of mentions; thus, in such a setting, the gold-standard coreference chain consist of a few large clusters, and any clustering which also merges most clusters will achieve high scores, even when many of the local linking decisions are wrong.

Luo *et al.* (2004) mention that a system which merges all mentions on the MUC-6 corpus will get 88.2% F-measure (100% recall and 78.9% precision), while realistic coreference systems get 81.9% F-measure (Harabagiu *et al.*, 2001), or 82.9% F-measure (Luo *et al.*, 2004), respectively. Luo *et al.* argue that the CEAF measure would avoid these problems as it is less lenient against the over-merging of coreference sets.

However, the problems created by evaluating on gold-standard mentions run deeper: Consider the results of Klenner and Ailloud (2008), who report results both for "true mentions" (i.e., only noun phrases from the annotation key) and all mentions. Even though using ECM/CEAF, an alignment-based metric that Luo (2005) claims to be less problematic for the evaluation on gold mentions (see subsection 3.2.3 for a discussion of alignment-based metrics), the evaluation results differ by as much as 13-18% between the two conditions despite the fact that treebank markables are used in both cases.

3.3 Summary

In this chapter, three aspects were covered that are central for understanding published evaluation results in coreference resolution: the first part covered annotation schemes and corpora annotated with these annotation schemes, including the two corpora most used in research for English – the MUC-6/MUC-7 and ACE coreference corpora – and the TüBa-D/Z corpus, which was used in the experiments that are described in the later chapters of this thesis.

Besides annotation schemes, and variations in their definitions of mentions, and which related phenomena to include in coreference resolution, the two other central aspects include the evaluation metric used on one hand, and the exact evaluation setting, on the other hand. Both of these aspects are important, and in the worst case – namely, the MUC evaluation measure together with a gold-mentions setting, no useful coreference resolver can outperform a trivial baseline (namely, creating one large coreference chain containing all mentions). Other evaluation metrics presented in this chapter include B-cubed and the ECM/CEAF-measure, which are the other common evaluation metrics that are often used in the literature. These measures are more sensitive to over-merging and are therefore usable in a greater range of settings. However, the exact evaluation setting is still important, as it has a considerable influence on the actual evaluation results.

Evaluations in "*true mentions*" settings (*no singletons*, *gold-standard subset*, *disregard spurious links*) or those assuming *perfect anaphoricity information* have a ratio of non-links to links that vastly differs from that occurring in a setting where the full set of noun phrases is used with no additional information. In more concrete terms, such a setting assumes or comes very near to gold-standard information on the discourse-newness of the anaphor (in the *no singletons* setting, it is known at least that the mention under consideration is not a singleton, and that none of the possible antecedents are singletons; in the *disregard spurious links* setting, incor-

rect links to singletons or from singletons are not counted, and in the *gold-standard subset* setting, the number of singletons is at least substantially reduced). "*True mentions*" settings also reduce the proportion of incorrect antecedent candidates, which means that high-recall methods are more likely to improve results than in a realistic setting.

Assuming gold-standard information on the discourse-newness of a potentially anaphoric mention, or that singleton mentions are removed, are assumptions that may in fact be rather unrealistic: Recovering this information (e.g., anaphoricity of definite noun phrases) from text is quite difficult and state-of-the-art methods for this task do not reach near-perfect accuracy (see section 4.2.3).

Considering this difference, section 6 and 7 will present experimental results in two settings: one in which the discourse-old/discourse-new distinction is assumed to be known (*perfect anaphoricity information*), and a more realistic setting in which discourse-new detection and resolution are integrated.

For both evaluations, I will use a link-based measure, which intuitively corresponds to scoring single decisions of the coreference resolver and does not share the undesirable properties of the MUC score in high-recall scenarios. Although link-based scoring is not applicable to other decoding models such as entity-mention models (Luo *et al.*, 2004) or clustering models (Culotta *et al.*, 2007), the link-based measures are easily interpretable and have a close relationship to actual model behavior in the case of models which link anaphoric expressions to an antecedent. Due to the difference between assuming the discourse-old/discourse-new information as given and a more realistic setting, I will provide results in both settings in chapters 6 and 7.

Chapter 4

Approaches to Coreference Resolution

The goal of this chapter is to outline the design space for a coreference resolution system, starting from the discussion in chapter 3 which outlined how evaluation results with respect to different settings can (or, more often, cannot) be compared to each other.

The following sections will provide an overview of the state of the art in coreference resolution systems, starting from early rule-based approaches in section 4.1.1, the evolution of machine learning based approaches to the point where they became the de-facto state of the art, in section 4.1.2. Extensions of this model that address the most important shortcomings are addressed in the following section 4.2. In particular, models that choose between multiple antecedent candidates at once (4.2.1), those that extend the strictly local model of antecedent choice to consider the larger context in the document including complete coreference chains (subsection 4.2.2), and the question of recognizing discourse-new mentions (4.2.3).

Section 4.3 discusses coreference systems that use richer semantic representations to overcome some of the problems of the shallow system. In section 4.4, I discuss existing coreference resolution systems for German. The summary in section 4.5 concludes by an overview over the findings of the earlier sections.

Early approaches to reference resolution assumed a small world setting in which human knowledge about relevant objects can be completely described. As a consequence, many of the problems that need to be solved in larger domains are trivialized and early research emphasized the necessity (and importance) of using common sense knowledge to allow reference resolution.

For example, Charniak (1972) mentions the following sources for relevant constraints on coreference resolution:

- Descriptive information (i.e., the contents of mentions): *When we look for "the ball", we can exclude anything which is not labeled* `ball`.

Note that Charniak works on abstracted propositional representations and is able to assume that all balls that have been mentioned have a `ball` label in the database without treating synonymy or paraphrases in his system.

- Recency information (i.e., preferring recent antecedents)
- Syntax rules (an earlier version of the binding restrictions that Reinhart, 1976 later formulates in terms of syntactic domain and c-command relations)
- Selectional restrictions: In "*He told Bill about the riot*", it is unlikely that *he* refers to my dog Rover.
- Other contextual information, which can be the result of multiple steps of inference.

Charniak then focuses exclusively on the inferences that are necessary for correct resolution in difficult cases such as example (4.1):

(4.1) Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a top. "Don't do that" said Penny. "Jack has a top. He will make you take *it* back."

It refers not to the most recently mentioned top (the one Jack has), but the (hypothetical!) top that Janet wanted to buy. This shows that world knowledge tends to come into the picture in some cases and that a complete model even of reference resolution would make a theory of common-sense reasoning with reasonable coverage necessary.

Current approaches to large-domain coreference resolution would essentially agree with the list of influence factors that Charniak's work outlines. However, accurate treatment of the contents of the mentions – involving the descriptive information as well as selectional restrictions – is significantly more difficult (and therefore error-prone) in a larger domain.

Approaches such as the one by Charniak (1972) are based on **common sense reasoning** in the sense that they construct a representation that is as close as possible to the representation that a competent reader may have. Such approaches are able to correct false predictions (that are due to recency or other salience factors) using inferences in the conceptual representation built by the system; but they crucially depend on the assumption that linguistic constraints as well as relevant domain knowledge can be completely represented in the system.

In a system that relies on inferences being made, incorrect (or even incomplete) knowledge of the domain will lead to incomplete or implausible inferences being made. In a large domain, using plausibility as a primary resolution criterion would result in considerable brittleness: On one hand it is not possible to model the domain well enough that the system will not run into unforeseen cases; on the other hand a model that relies on being able to cover every case in depth will be prone to

catastrophic failure in such 'exceptional' cases that are not covered by the domain model.

Contributing to the attraction of such inference-based approaches is a tendency to over-emphasize common sense and plausibility as a constraint in the interpretation of a discourse. The interpretation is determined by a multitude of factors, of which plausibility (in a common sense reasoning way) is only one and may either conflict or even be overruled by linguistic preferences. For example, in (4.2), resolving *it* to *Windows 7* (and suggesting that Microsoft forces people to stop using Windows 7 when it comes out) is implausible. Despite this fact, the alternative resolution (to *[Windows] XP*) is not immediately available. Similarly, in (4.3), the reading where the pronoun '*it*' is resolved to '*your baby*' is readily available despite a dispreference for boiling babies (even in linguistic examples) and a dispreference to use *it* to refer to animated entities, which indicates a strong syntactically-mediated preference in this case.

(4.2) After Windows 7 comes out in October, will Microsoft somehow force us XP users to stop using [it]?

(4.3) If your baby does not thrive on raw milk, boil [it].

Compare this to the example (4.4), taken from Sidner (1979), where a strong enough selectional restriction (as opposed to a mere preference) blocks *a big graduation party* as antecedent for *it*:

(4.4) [$_1$ Cathy] wants to have [$_2$ a big graduation party] at [$_3$ her house].
[$_1$ She] cleaned [$_{3/*2}$ it] up so there will be room for everyone who's coming.

In most naturally occurring discourse, interpretation is over-determined by both plausibility restrictions and linguistic restrictions and preferences, which means that an over-reliance on common sense knowledge to explain the selection of the correct interpretation may be prone to explain away other influences which might actually be more tractably modeled for larger domains.

In summary, there are cases where either semantic or general world knowledge is necessary, and where representing (approximated) knowledge is necessary to improve the coverage over that of shallow systems. Thus, research in the long term will (eventually) have to find ways to deal with logical relations between mention contents (Charniak's *descriptive information*), selectional restrictions, as well as discourse-based restrictions such as those posited by Asher and Lascarides (1998). In doing so, models for coreference resolution on small domains may help us to pinpoint the types of knowledge involved and the ways this knowledge is used, but it will be necessary to find approaches where imperfect information does not render the system brittle in a larger domain.

4.1 Towards Modern Coreference Resolution

In the following two subsections, I will outline the work that led to the systems of Soon *et al.* (2001) and Ng and Cardie (2002c), which are often used as the blueprint of a 'standard' coreference system (and often reimplemented to provide an intelligent baseline).

The contribution of subsequent research can usually be seen in one of two large groups: One (discussed in 4.2) is research that aims at finding a better strategy for coreference resolution than the simplistic "go back until the classifier says yes" strategy of Soon *et al.* (explained in more detail in section 4.1.2), including anaphoricity detection, approaches that improve the antecedent selection mechanism by comparing multiple plausible antecedents, and approaches that enforce global coherence (e.g., with respect to gender) in the resulting coreference chains.

The other group (discussed in section 4.3) consists in research that includes more informative features than in the basic Soon *et al.* model, aiming at improving name matching, but also at resolving some of the cases which we earlier subsumed under 'coreferent bridging', where surface cues are no longer enough as a robust indication of coreference.

The fact that Soon et al.'s approach is often taken as a baseline, however, should not be taken as an indication that it is somehow groundbreaking in the behavior of the system: while often not described in adequate detail, and as a result rarely cited by modern work, rule-based approaches did use the same or an even greater variety of information sources than Soon *et al.* and their successors, and achieved results in the earlier MUC evaluation that still provide a strong baseline.

Table 4.1 summarizes performance results on the MUC-6 and MUC-7 datasets. The table contains some counterintuitive results: While other fields such as parsing, have seen statistical models quickly overtake their hand-crafted counterparts, classifier-based coreference resolution brought no such landslide.

The first successful machine-learning model of Soon *et al.* – with a distance of several years to earlier machine learning models such as the one by McCarthy and Lehnert (1995) – is visibly behind the most successful rule-based models of Kameyama (1997) or Humphreys *et al.* (1998). And while it would be intuitively plausible, and indeed desirable, that more ambitious proposals bring drastic improvements, care has to be taken not to count as a drastic improvement what is actually just an artefact of the chosen non-standard evaluation setting. For example, complex path search in WordNet, paired with an elaborate learning approach as proposed by Harabagiu *et al.* (2001), or Bayesian unsupervised learning (Haghigi and Klein, 2007) represent approaches to tackle the problem of low recall in coreference resolution. The evaluation scores published in these papers are higher than for earlier, more pedestrian approaches; however, the results of Harabagiu *et al.* and of Haghigi *et al.* are derived in a simplified setting (called *No Singletons* in the earlier discussion) that makes evaluation results incomparable to the actual coreference task.

MUC-6	MUC-F
(Kameyama, 1997)[a]	64.8
(Lin, 1995)[a]	63.0
(Fisher *et al.*, 1996)[a]	47.2
(Soon *et al.*, 2001)	62.6
(Ng and Cardie, 2002c)	70.4
MUC-7	**MUC-F**
(Humphreys *et al.*, 1998)[a]	61.8
(Lin, 1998d)[a]	61.1
(Soon *et al.*, 2001)	60.4
(Ng and Cardie, 2002c)	63.4
(Uryupina, 2006)	65.4
MUC-6 ("true" mentions)	**MUC-F**
baseline: merge everything	88.2
(Harabagiu *et al.*, 2001)	81.9
(Luo *et al.*, 2004)	83.9
(Luo, 2005)	90.2
(Haghigi and Klein, 2007)[b]	70.3

Results ordered chronologically, with original participants first independently of publication year.
[a]: participants of the original shared tasks (no use of testing data for development) [b] unsupervised learning using a larger quantity of gold-mention information without coreference chains
MUC-F: F-measure in the model-theoretic scoring scheme of Vilain *et al.* (1995).

Table 4.1: Evaluation Results reported on MUC-6/MUC-7

ACE-02 (system mentions)	MUC-F	CEAF	ACE-val
(Luo *et al.*, 2004)			73.4
(Daumé III and Marcu, 2005)			79.4
(Yang and Su, 2007)[a]	62.7–67.1		
ACE-02 ("true" mentions)	**MUC-F**	**CEAF**	**ACE-val**
(Luo *et al.*, 2004)		73.2	89.8
—, only string matching / alias		64.4	86.0
(Daumé III and Marcu, 2005)			89.1
(Ji *et al.*, 2005)[b]	83.7		
(Ponzetto and Strube, 2006)[c]	70.7		
(Ng, 2007)[d]	64.5	62.3	
(Denis and Baldridge, 2007a)[e]	67.5–72.5		
(Haghigi and Klein, 2007)[f]	62.3–64.2		

[a] Yang and Su only provide separate scores for different sections of the ACE-02 data set. [b] Ji *et al.* use a setting where links between non-key mentions are discarded. [c] Ponzetto and Strube use a setting where system mentions are aligned with the gold standard and non-aligned mentions are discarded. [d] Ng does not mention the exact setting in his paper; other papers suggest that his system delivers similar results for both gold-standard and system-generated mentions. [e] Denis and Baldridge only provide separate scores for different sections of the ACE-02 data set. [f] Haghighi and Klein report results on a subset of the training sets for an unsupervised approach.

MUC-F: F-measure in the model-theoretic scoring scheme of Vilain *et al.* (1995); CEAF: ECM/CEAF measure as in (Luo, 2005); ACE-val: official ACE evaluation measure.

Table 4.2: Evaluation Results reported on ACE-02

Furthermore, as Luo *et al.* (2004) point out, this latter evaluation setting results in a rather high evaluation score of F=88.2% for a trivial baseline where all (non-singleton) mentions are grouped in one large coreference chain, which no system can improve on unless it over-merges drastically enough that it would be useless for any real-world purpose (Luo, 2005). In the more realistic setting with system mentions, Uryupina (2006) uses a feature-rich coreference resolver using flexible string matching, syntactic information and WordNet similarity and achieves a visible improvement which is however small in comparison to the results claimed in the *No Singletons* setting.

The situation of results reported on the ACE corpora is slightly more complicated than on the MUC corpora, as various researchers choose their evaluation measure differently from the ACE-value of the official ACE scorer, the MUC metric and different other proposed metrics such as ECM/CEAF or Bcubed. Once again, results on system mentions and on "true" mentions can hardly be compared; the best system on automatically generated mentions according to ACE-value, by Daumé III and Marcu (2005), with an ACE-value of 79.4, lies below the ACE-value for a simple string matching/alias baseline in the *Gold Markables* setting. Some researchers such as Ponzetto and Strube (2006) and Ji *et al.* (2005) choose more

extreme simplifications of the setting (Ponzetto and Strube eliminate all singletons, whereas Ji *et al.* additionally remove all non-identified gold markables from the answer key), with many researchers not clarifying the exact evaluation setting used.

Nonetheless, it is desirable to make progress past the status quo of classifier-based, knowledge-poor coreference resolution even if it turns out to be substatially harder than could be expected from only looking at results from "true mentions" settings. That doing so is possible is demonstrated by research such as that by Uryupina (2006) and Yang and Su (2007), which achieves improvements in a realistic setting.[1]

4.1.1 Rule-based approaches to Coreference Resolution

Most participants in MUC relied on relatively simple heuristics to match names and variations thereof, in addition to heuristics that allowed to resolve certain cases of anaphoric pronouns.

Other researchers used relatively elaborate mechanisms: as an example of a coreference resolution component that does not use machine learning, we'll use the coreference component of the PIE system (Lin, 1995), that integrates information from different sources to arrive at a clustering of the mentions: basic rules either indicate coreference of two mentions or their non-coreference, and constraints are applied by decreasing priority, whereas lower-priority constraints that would conflict with constraints imposed by higher-priority rules are ignored. Constraints can either have a positive weight to indicate that the two mentions should be merged, or a negative one, to indicate that two mentions should be marked as incompatible and the clusters they are part of never should be merged.

Lin's constraint propagation algorithm works by ordering the constraints by their absolute value, and then either merging two mention clusters (for a positive value), or adding an inequality constraint. Subsequently, lower-ranked constraints that have become redundant (when they predict the coreference of two mentions that are already in the same cluster) or conflict with the decisions already taken are discarded, and the process is repeated until no constraints remain.

PIE uses name matching and a unique names assumption (saying that different names have to refer to different entities) for names, binding theory restrictions as well as preferences derived from Centering Theory for pronouns, and semantic/syntactic compatibility informed by WordNet for common nouns, where positive weights (i.e., indications of coreference) are adjusted according to the proximity of the two noun phrases and the grammatical role of the antecedent.

Kameyama (1997) discusses the resolution algorithm that was developed for SRI's FASTUS system (Hobbs *et al.*, 1996), which also participated in MUC-6, and

[1] One should also mention the work of Daumé III and Marcu (2005) here, where the authors make a point of integrating ACE mention tagging and coreference resolution, and indeed create a very well-performing coreference system, but then proceed to discuss the usefulness of features based on the gold-mentions setting.

scored a recall of 59% and precision of 72% (i.e., F=0.65) in the blind evaluation.

Kameyama uses FASTUS' rule mechanism, an extended finite state calculus which allows extraction of text parts, to create a record for each nominal expression with slots fillers from the text including information about the determiner type, grammatical number, the head string with possible sorts from a sort hierarchy, and modifier strings.

In a subsequent step, possible antecedents are collected for each potentially anaphoric mention in the sentence, using the whole text for mentions in the `HEADLINE` region, or preceding text in the `TEXT` region (including the whole text for proper names, ten sentences for definite descriptions, three sentences for pronouns, and only the current sentence for reflexives). Intrasentential cataphoric antecedents are collected for first-person pronouns.

Possible antecedents are then filtered to enforce number consistency, as well as enforce semantic consistency by requiring that the anaphor has the same or a more specific sort[2] than the antecedent candidate, and by filtering out pairs with incompatible modifiers (e.g., *French* and *British*). The subsequent ranking uses a left-to-right order in the same sentence as the mention to be resolved (approximating syntactic preferences as they are manifest in Hobbs' 1978 algorithm), then potential candidates in left-to right order in the preceding sentence, and a right-to-left order in the other sentences.

In addition to this, Kameyama's resolver contains a special-purpose alias recognition algorithm that captures name variations (e.g. *Colonial* for *Colonial Beef*) and an acronym detector that matches a selective sequence of initial characters in the full name (e.g. *GM* for *General Motors*).

4.1.2 Introducing Machine Learning

The view of coreference resolution as a task of information fusion, where several information sources are combined to provide more definitive information, begs for the application of machine learning techniques, as these are usually faster and/or more efficient than humans at finding effective combinations of features.

One such approach is described by Burger and Conolly (1992), where evidence is combined in a hand-crafted Bayes network. The Bayes network is meant to faithfully model inter-relationship between several variables that are predicted by theoretical accounts of anaphoric reference (such as agreement or recency). Although no formal evaluation is attempted, the paper puts forward the idea that "probabilistic treatment more accurately models the 'non-deterministic' way in which linguistic evidence supports reference hypotheses".

Later work (Connolly *et al.*, 1994) abandons the generative idea and focuses on learning a decision function which can be used in the process of providing an

[2] The sort hierarchy is not detailed in Kameyama's or any other paper on the FASTUS system as used for MUC-6, but the description seems to indicate that it is a shallow taxonomy that covers the most important types of the MUC domain, and contains concepts such as automaker, airline and company.

antecedent. In their model, a machine learning classifier is learnt for deciding between two antecedent candidates, given an anaphoric definite or pronoun. Then, starting from the two farthest candidates, the classifier is required to choose between the two to update the current 'winner' hypothesis, and the process is repeated by comparing the next candidate with the current best hypothesis.[3] As features for the classifier decision, they use gender and number agreement, recency, grammatical roles of anaphor and candidates, and the mention types of anaphor and candidates, in addition to subsumption in a manually constructed system for knowledge representation. They evaluated this algorithm in conjunction with several learning methods, as well as hand-crafted decision lists for pronouns and definite descriptions. Several learners achieved performance on par with the hand-constructed classifier or even surpassing it, including a statistical back-off model, decision trees and subspace-trained neural networks (using a variant of backpropagation training where features are partitioned into interdependent subsets and training is performed with these parts of the features separately). The best performance is attained by the subspace-trained neural network that finds the correct antecedent for 37.4 percent of definite descriptions and 55.3 percent of pronouns.

A trend that can be observed here is the use of *off-the-shelf* machine learning techniques: instead of investing a lot of work into a tight coupling of learning techniques and features, as with the manually constructed Bayes networks, the machine learner is used to construct a decision function, with features as input and the decision as output. This engineering decision makes it possible to focus on the construction of features by treating the machine learner as a *black box*, but at the same time makes it necessary to consider the *mapping between classifier decisions and system output*. In Connolly et al.'s system, the classifier is trained with pairs composed of the correct candidate with other potential (but incorrect) candidates, and testing was done in an elimination process where the current hypothesis is compared in turn with each other potential candidate, retaining either of both as the new current hypothesis based on the classifier decision. Another aspect worth noting is that Connolly et al.'s resolver only *resolves* anaphoric definite descriptions and pronouns, but makes no attempt at finding out whether a definite description has an antecedent or not.

McCarthy and Lehnert (1995) developed a decision-tree-based coreference resolver, later called RESOLVE, for the use in the MUC-5 information extraction task. In RESOLVE, a decision-tree classifier is learnt on all pairs of mentions from the document – in the context of MUC-5, all pairs of noun phrases relevant to the template-extraction task – with a desired output of whether the two mentions of a pair are coreferent or not. In testing, the classifier is applied to all pairs of mentions, and mentions are linked into a coreference chain if the classification is positive. (Note that this strategy does not require all pairs in a coreference chain

[3] Starting from the farthest mention means that more recent antecedent candidates have to 'win' fewer comparisons than farther antecedent candidates in order to be selected as the best-hypothesis antecedent.

to have a positive classification, as they can also be implicitly linked via multiple 'explicit' links formed by positive classifications).

McCarthy and Lehnert used both domain-independent features (mention type, name substring, both mentions of a pair being in a common noun phrase, or both mentions being in the same sentence) and domain-specific features (either or both mentions of a pair being a company created in a joint venture) as input to the decision tree. An evaluation on hand-annotated texts from MUC-5 shows that the learnt decision tree gives much better recall than the previous rule-based system (described in Lehnert *et al.*, 1992) at a very small cost in precision.

A later version of RESOLVE was evaluated in the MUC-6 coreference task (Fisher *et al.*, 1996), using features such as the two mentions having the same string, the earlier mention being the most recent compatible subject, having the same semantic type, or one or both of the mentions being a role description for a person (using several patterns to match common forms of role descriptions). Blind evaluation results for MUC-6 are 44% recall and 51% precision, a rather low result compared to the performance on the development set, but also in comparison to the best-performing systems in MUC-6 such as the rule-based system by Kameyama (1997), something that Fisher et al. attributed to the focus on person and organization mentions,[4] as well the decision to use an unpruned decision tree – decisions which turned out to worsen the performance on the official evaluation data.

The system of Aone and Bennett (1995) resolves both anaphoric pronouns, anaphoric definite noun phrases and name coreference in Japanese text using a machine-learning approach based on decision trees. Zero pronouns, anaphoric definites and discourse-old name mentions are marked up in the text by hand, and are resolved by pairing the anaphoric expression with each possible antecedent, creating *feature vectors* from information that was extracted by syntactic preprocessing. These feature vectors are then classified by a decision tree learned earlier, and among the possible antecedents that are classified as positive (i.e., coreferent), the one with the highest confidence is taken, breaking ties between multiple antecedents with highest confidence values by selecting the closest antecedent.

The feature set of Aone and Bennett's system is based on an earlier rule-based system and includes distinctions such as semantic classes of anaphor and antecedent candidate, NP form and semantic class of the involved noun phrases.

To train the decision tree classifier, an anaphor in the training data is paired with all prior members of its coreference chain, and using the feature vectors created in this way as positive examples, while negative examples are created by pairing an anaphor with all previous mentions that are not coreferent.

Soon *et al.* (1999, 2001) present a decision-tree-based system for coreference resolution that is evaluated on the MUC6 and MUC7 corpora. In contrast to McCarthy and Lehnert (1995) and Aone and Bennett (1995), they explicitly assess the influence of preprocessing, saying that their preprocessing recovers 85% of the

[4] According to McCarthy (1996), only about 50% of the mentions in MUC6 referred to persons or organizations.

mentions in MUC6, and that they do not focus on a small number of semantic classes that are central to an information extraction task (McCarthy and Lehnert, as well as Aone and Bennett, focus mostly on organizations and persons and do not even attempt to find coreference relations for other mentions).

Soon *et al.* use a number of features, including some that are indicative of the form of the noun phrase (the anaphor/antecedent being a pronoun, the anaphor being a definite/demonstrative noun phrase, as well as a feature indicating that both mentions are proper names), agreement in number, gender, or semantic class, distance in sentences between the two mentions, as well as string matching (stripping off determiners, i.e., essentially comparing the head and the premodifiers) and an alias feature that copes with name variation (allowing the matching of "Mr. Simpson" and "Bent Simpson"), common organizational suffixes, and abbreviations.

With the decision tree learned in MUC6, the system effectively resolves to the closest mention that either:

1. has a matching string (string matching feature)
2. is detected as a name alias
3. is a gender-matching antecedent in the same sentence for a pronoun anaphor.

The fact that this system does better than a majority of the participants in the MUC6 evaluation despite the simplicity of the feature combination is rather surprising. The explanation of this might be that, on one side, using machine learning to handle the combination of the features allowed Soon *et al.* to put more work into the preprocessing pipeline, but also to avoid mistakes such as only resolving persons and organizations (which was one of the factors responsible for the suboptimal performance of RESOLVE in the submission by McCarthy and Lehnert, 1995).

On the other hand, considering Kameyama's (1997) analysis, the system successfully reaps a number of low-hanging fruits in the textual domain used for MUC: it resolves same-sentence pronouns, which according to Kameyama constitute a sizable part of the pronouns in MUC6 and are relatively easy to resolve, and it also does a relatively good job at resolving name coreference. At the same time, the string matching feature allows to resolve easy cases of common noun phrases (i.e., those that have the same set of premodifiers).

Ng and Cardie (2002c) present a variation of the Soon *et al.* method, where they start from a system that uses the same features as Soon et al., but with improved preprocessing – their preprocessing recovers more of the key markables and the whole system then achieves 66.3% F-measure instead of Soon et al.'s original 62.6%, or 61.2% on MUC7 instead of Soon et al.'s 60.4%. They then modify the framework in three different points:

- In training, cases where the previous element of the coreference chain of a full noun phrase is a pronoun are treated differently: Ng and Cardie use the previous full noun phrase of that coreference chain, and generate negative instances from non-pronominal mentions up to that antecedent.

- Instead of the closest mention that gets classified positively by the classifier, they use the mention that the classifier has positively classified with the greatest confidence (defined as an add-1-smoothed probability).
- Instead of a single string match feature, they use three separate features for pronouns, proper names, and common noun phrases.

Together, these modifications give an additional 1.2% improvement (i.e., 67.5 for the system using different string matching techniques, training instance selection, and inference technique instead of 66.3), mostly through increased precision. They also add additional features: out of an initial set of 53 features that proved to be too much for a decision tree classifier to handle, they selected 26 by eliminating all features that they found to lead to low-precision rules. The decision tree that is then learnt on MUC6 classifies a pair of mentions as positive when:

- one is an alias of another
- they are common nouns with matching strings and their animacy could be determined.
- they are pronouns with matching strings
- the antecedent candidate is the preferred antecedent of an external pronoun resolution algorithm and not a premodifying bare noun, and it is either a pronoun with matching animacy value or a non-pronoun with a positive value for semantic class compatibility and where both NPs have the same maximal NP projection.
- They are identified as appositives and the relation is not ruled out by gender incompatibility.

With the hand-selected features, the F-measure increases by another 1.6% (from 67.5% to 69.1%) on MUC6; performance on MUC-7 increases by 0.4%.

4.2 Refining Machine Learning Approaches

The closest-first approach of Soon *et al.* shows promising evaluation results, and, as a machine learning approach, it should in principle be extensible by simply adding additional features. Yet, as witnessed by Ng and Cardie in their work, the approach forces the user to second-guess the machine learner and perform manual feature selection (or, only slightly better, use wrapper-based feature selection methods on a development set) to incorporate new features, and makes it necessary to use non-trivial strategies for sample selection (see for example Ng and Cardie 2002a). This should be seen as an indication that Soon et al.'s method and its variants are only moderately successful at relieving the developer from having to think about nontrivial decoding issues beyond the features themselves and only partly fulfills the promise of an easier life by using machine learning methods.

The question, then, is how a machine-learning based coreference system should look if the goal is to incorporate information about salience, compatibility, anaphoricity by means of new features.

One perspective on this problem would be to stick with the view that a coreference resolution system looks at all the mentions in the text (in reading order) and tries to resolve any anaphoric expressions it finds (or more generally, find an antecedent for any discourse-old mention it encounters).

In this perspective, coreference resolution consists in (i) identifying anaphoric expressions (pronouns or definite noun phrases) and (ii) resolving them. As we will see, it is not always clear how to separate these two subtasks; in fact, the model by Soon *et al.* does not even attempt to separate them. Ng and Cardie (2002b) train a classifier for anaphoricity of noun phrases and observe that it only helps performance if the anaphoricity classifier is overridden for cases where string matching and alias detection find an antecedent.

A different perspective on this problem is a more global view where coreference is seen as a clustering problem: since the goal for coreference resolution is a partition of the mentions according to coreference, algorithms and descriptive terminology from clustering can be readily applied. The clustering metaphor works comparably well for the related task of cross-document coreference (Bagga and Baldwin, 1998b; Mann and Yarowsky, 2003), and correspondingly for the coreference resolution of names. However, the clustering metaphor is not very helpful for cases where coreference links are necessarily asymmetric and local (for example in the case of pronouns, where an antecedent usually is local, but further elements in the coreference chain do not have to be close to the anaphoric mention).

4.2.1 Approaches to Antecedent Selection

With respect to the selection of possible antecedents, the system of Soon *et al.* (2001) always selects the closest candidate, a strategy that does reflect some approximate salience information (i.e., closer candidates being preferred), but does not leave any way to influence that ranking. Ng and Cardie (2002c) do address this, and resolve to the antecedent that the classifier identified with the greatest certainty. Research on mapping multiclass problems to binary ones in machine learning suggests that this strategy does not always lead to a good classifier: Beygelzimer *et al.* (2008) report that there is no known bound for the regret (the difference between the obtained classifier and the best possible multi-class classifier) when converting weighted one-against-all classifications from optimal binary classifiers to a multiclass classification in the presence of noise (i.e., if the data is not perfectly separable).

Tournament models Yang *et al.* (2003) use an all-pairs tournament to choose among candidates. In training, a classifier with features pertaining to two different antecedent candidates is trained with examples where one of the antecedent candidate is the correct one and the other one is an incorrect one. (The classification

is whether the more recent one or the more distant one is the correct one). In testing, all possible candidates are paired with each other and for each time that the twin-candidate classifier chooses a given candidate, the score of that candidate is incremented; the antecedent candidate with the highest score is selected.

The candidates are selected using the following approach: For pronouns, a candidate is chosen among all noun phrases in the last two sentences that do not disagree in number, gender, or person (going further back if no compatible candidate can be found). Antecedent candidates for non-pronouns are selected as in (Ng and Cardie, 2002c) by choosing those instances that are classified as positive by a single-candidate classifier, whereas training data for these is created by using non-pronominal antecedents and adding negative examples from the surrounding text (previous, current and next sentence seen from a possible antecedent). Their approach outperforms the single-elimination tournament of Connolly *et al.* (1994) by a large margin and also gives higher performance than the one-against-all approach of Ng and Cardie (2002c).

Direct ranking Machine learning methods that allow a classifier to choose among an open set of candidates have become available only recently; most of the time, the algorithm is at first only formulated for the case of binary classification or for selecting one out of a fixed set of labels. On the other hand, efficient algorithms for ranking problems (i.e., choosing from an open set) are available and have been used in other fields such as parsing (Johnson *et al.*, 1999) or language modeling for speech recognition (Rosenfeld, 1996). The direct use of ranking classifiers has been investigated by Versley (2006) for coreferent bridging and by Denis and Baldridge (2007b) for pronoun resolution. The findings both of Versley (2006) and of Denis and Baldridge (2007b) indicate that ranking classifiers - for these tasks - do perform better than one-against-all or all-against-all tournament classifiers.

4.2.2 Global Models of Coreference

All the models mentioned before make strictly local decisions. Not only is it impossible to take into account salience information related to the number of previous mentions, it is even possible to end up with a non-consistent chain where a mention of “Mr. Clinton” is linked by a mention of “Clinton” and then “she”, regardless of the fact that “Mr. Clinton” clearly has male gender and “she” clearly has female gender.

Constraint propagation To enforce the global consistency of coreference chains, it is necessary to look beyond single links and consider (partial) coreference chains, weighing positive and negative evidence against each other. While machine learning is not a necessary part in such approaches (witness Lin’s 1995 PIE system mentioned earlier), the complex interaction of rules or information sources in such approaches is cumbersome to manage by hand, and thus methods that can learn parameters of a system are highly attractive.

Harabagiu *et al.* (2001) present a model that contains both hand-written and learned rules over anaphor and antecedent, such as "*when the anaphor is a pronoun, and has the same string as the antecedent, link the anaphor to the antecedent*".

They determine rule weights by counting correct and incorrect cases where the rule may apply, and give a weight to the rule that depends on the proportion $p^+(R_i)$ of correct cases to all cases resolved. The weight is then determined as a nonlinear function of $p^+(R_i)$, using the entropy of the distribution of positive/negative cases:

$$rel(R_i) = \begin{cases} 1 - entropy(R_i) & \text{if } p \geq n \\ -(1 - entropy(R_i)) & \text{otherwise} \end{cases}$$

When testing, they perform agglomerative clustering to maximize an objective function over partitions Par using within-cluster and between-cluster similarity:

$$\sum_{(\mathrm{NP}_i,\mathrm{NP}_j)\in Par} score(\mathrm{NP}_i,\mathrm{NP}_j) - \sum_{(\mathrm{NP}_i,\mathrm{NP}_j)\notin Par} score(\mathrm{NP}_i,\mathrm{NP}_j)$$

where $score(\mathrm{NP}_i,\mathrm{NP}_j) = \sum_{R(\mathrm{NP}_i,\mathrm{NP}_j)} rel(R)$ uses the rule weights learned in training to give positive or negative weights to single links. To avoid local minima, the clustering procedure does not proceed greedily, but considers the five highest-scoring partitions found for further merging.

Harabagiu *et al.* use this approach together with hand-crafted rules and semantic information from WordNet with the entropy-based weighting discussed above. Since Harabagiu *et al.* assume perfect discourse-new information, their evaluation results are hardly comparable to other work (see table 4.1: Harabagiu's results are at a MUC F-measure of 81.9% , which is visibly below a trivial baseline of merging all mentions into one cluster, which would yield 100% recall and 78.9% precision, or 88.2% F-measure in the scoring scheme of Vilain *et al.*, 1995 – but they are also much higher than any result that has been achieved in a realistic setting).

Entity-mention models Yang *et al.* (2004) capitalize on the insight that earlier elements in the coreference chain may contain discriminating information by including several features dependent on previous coreference decisions in the feature set used for their link-based classifier. While their classifier makes decisions strictly locally (i.e., in a greedy fashion), this allows them to avoid some disagreements in number, gender, semantic class and use the NP with the longest string for matching:

If we have a partial chain including "Mr. Clinton" and "Clinton", the chain-based features would keep the system from linking "she" to "Clinton" as the latter is in a chain with the attribute "male". However, strong evidence that "Clinton" and "she" corefer would not lead to the system reverting its decision to link "Mr. Clinton" and "Clinton".

On a corpus of coreferentially annotated medical abstracts, the system including cluster-based features achieves a visible improvement in both precision and

recall over a Soon et al.-style baseline system. A more elaborate version of this approach is presented in Yang *et al.* (2005a), where the existence of an NP in the antecedent cluster with a matching string is introduced as a feature. The approach is shown to yield an improvement over both a Soon et al.-style baseline as well as best-first clustering in the style of Ng and Cardie. Yang *et al.* also discuss a variant they call 'greedy' clustering, where several rounds of cluster-merging are undertaken, while every cluster is still represented by its leftmost member for the purposes of classification. The 'greedy' clustering does not yield any obvious improvement over the incremental strategy.

Beam search for entity-mention models Daumé III and Marcu (2005) present a model using global features not unlike the model of Yang *et al.* (2004). In contrast to Yang *et al.*, they complement the non-local features with a machine learning model that is able to avoid decisions that later turn out to be non-optimal by keeping track of multiple alternative solutions (*beam search*). Daumé and Marcu's system chooses antecedents by means of online learning, using a large-margin variant of the perceptron algorithm (Gentile, 2001), which allows them to explore a larger space of linking strategies without having to patch up problems in sample selection. They use a strategy they call *intelligent link*, with the following linking approach:

- Names are first matched to other names in the previous document, otherwise, against the last nominal, or, failing that, using the highest-scored link.
- Nominals are matched to the highest-scoring nominal in the previous chain, otherwise, against the most recent name; failing that, using the highest-scored link.
- Pronouns are resolved with an *average-link* approach against all pronouns or names, or, failing that, using the highest-scored link.

Linking pronouns only to pronouns or names makes sense in an ACE-type scenario (where only coreference chains for referents of a small set of classes are wanted), since most pronouns will either refer to persons or organizations (which are usually named), or to non-ACE mentions (in which case they do not have to be linked). Using clusters also allows Daumé III and Marcu to include a "decayed density" for each entity, somewhat similar to Lappin and Leass' (1994) salience measure, exploiting the fact that some entities are very central to a document and are referred to throughout the whole document, whereas other cases of pronominal coreference are only very local.

Daumé III and Marcu reach an ACE coreference score of 79.4 for joint mention detection and coreference resolution, against 78.1 for performing these two steps in isolation; using perfect mentions, they get an ACE score of 89.1.[5] Compared with the original participants of the ACE-2004 evaluation, Daumé III and

[5]The ACE score is a performance measure developed for the ACE evaluation that does not directly correspond to any intuitively plausible figure of merit. Different ACE evaluations used slightly different variants of the ACE score. (David Day, *p.c.*)

Marcu's system lies between the second-placed entries for joint mention detection and coreference resolution, and between the first-placed and second-placed entries for coreference resolution on gold-standard markables.

Supervised clustering Culotta *et al.* (2007) represent coreference resolution as a clustering task. Casting the decoding process as agglomerative clustering (i.e., eliminating local decisions as in the local models or Daumé III and Marcu's global model altogether), clustering possibilities are evaluated by a linear combination of pairwise features, which are not unlike the ones used by earlier approaches and set-based features that can help enforce consistency constraints (such as consistency of number and gender across one cluster) and can also help exclude nonsensical possibilities such as clusters consisting of only pronouns. Culotta *et al.* use the results of an external pronoun resolver as a feature, which helps avoid the loss of performance that is incurred when abstracting from mention pairs in other supervised clustering models such as Finley and Joachims (2005).

4.2.3 Anaphoricity determination

As mentioned earlier, not all noun phrases we examine are discourse-old. In fact, even of the definite common noun phrases, less than half refer to an entity that has already been mentioned in the previous discourse.

Syntactic heuristics Vieira and Poesio (1996) used several heuristics for filtering out discourse-new noun phrases in the VPC corpus, which is based on text and syntactic annotation from the Wall Street Journal:

- Proper nouns are generally taken as discourse-new if they have not occurred in the previous discourse.
- Certain noun phrases are used very commonly to refer to larger-situation or unfamiliar referents, especially time-related terms (*year, day, week, month, hour, time, morning, afternoon, night, period, quarter*), terms that denote abstract objects (*fact, result, conclusion*) or modifiers that create a functional expression (in the sense of Löbner, 1985) out of a predicate, such as *first, last, best, most, maximum, minimum, only, more, closer, greater, bigger* and other superlatives.
- Definites that occur in an apposition position or in certain copular constructions do not normally refer to previously introduced referents. An exception for this are definite noun phrases occuring in a copular construction with an adjective.
- Vieira and Poesio found that postmodification (by prepositional phrases or relative clauses) almost always indicates a discourse-new referent.

Vieira and Poesio find that their system achieves R=66.3% and P=84.6% for discourse-new mentions on their corpus.

Unsupervised learning Bean and Riloff (1999) present work on the same problem, on the basis of news text from the MUC-4 corpus. In addition to syntactic heuristics, they used unsupervised learning to acquire discourse-new information from an unannotated corpus. The following sources of information are used:

- The definite-only (DO) information source simply extracts all noun phrases that occur at least 5 times and always carry a definite article, extracting noun phrases like *the contrary* or *the National Guard*.
- Definite NPs occuring in the first sentence of a text (S1) and not classified by the syntactic heuristics. Since non-unique (or, in the terminology of Bean and Riloff, non-existential) noun phrases can be felicitously introduced as discourse-new when they occur with restrictive modification, this is especially important. (For example, "*the president of the United States*" is a unique description, whereas "*the president*" is non-unique and should not be put on the list when the former noun phrase occurs in the first sentence).
- Bean and Riloff also try to generalize the noun phrases coming from the S1 extractor by constructing patterns from them. Starting from the head noun, they try to progressively expand the pattern with additional nouns to cover composite noun phrases.

  In this way, multiple noun phrases that are extracted by the S1 heuristic, such as "*The Salvadoran Government*" and "*The Guatemalan Government*" are abstracted into a pattern such as "*the* $\langle x^+ \rangle$ *Government*". Bean and Riloff use the term *existential head pattern* (EHP) for these patterns.

To eliminate false positives from the S1 and EHP list, they applied a threshold on the NPs' definite/indefinite ratio, such that NPs below a lower threshold are still classified as non-unique even when they are extracted by S1 or EHP, whereas NPs where the ratio is not below the lower threshold, but below an upper threshold are classified as *occasionally existential*. Occasionally existential noun phrases are not resolved if they occur within the first three sentences of a text, but are treated as discourse-old when their first occurrence is after the third sentence in the article.

Bean and Riloff report a 84.5% precision and 79.1% recall on *discourse-new* mentions[6] for the isolated task of anaphoricity detection. Bean (2004) reports results for the integration of the anaphoricity detector into a coreference resolver, with large gains in precision (for both definite descriptions and other anaphoric mentions), which is however offset by a large loss in recall, which means that the F-measure for definite descriptions does not improve at all (55% for definite noun

[6] Note that discourse-new mentions are in the majority, which means that recall and precision for the complementary class of discourse-old mention would be more interesting.

phrases in the terrorism domain, both with and without filtering, and 60% with and 59% without filtering on their natural disasters texts). Overall F-measure shows a slight improvement (60% to 61% on terrorism texts) or a slight degradation in performance (64% to 63% on natural disasters texts).

A discourse-new classifier Ng and Cardie (2002b) present work combining several heuristics with the help of a machine learning classifier and evaluate the use of their discourse-new classifier on the MUC-6/7 corpora.

Ng and Cardie used several syntactic cues not unlike those used by the earlier approaches:

- being in an appositive or predicate nominal construction
- postmodification
- presence and type of premodifiers of the mention head, such as proper names, or numbers occurring as premodifiers.
- presence of superlatives or comparatives

They also use a fine-grained classification of the mention (including information whether the mention has a definite/demonstrative/indefinite article or is quantified).

To identify potential previous mentions, they have indicators for string-matching and same-head candidates, but also for name variation and hyponyms that occur before the current sentence, as well as some zoning information (being entirely in uppercase; being in the body of the text; being part of the first sentence; being in the first paragraph). In addition, they use semantic categories typically used for referring to people (such as the mention being a post, or the title of a person).

The classifier based on this information achieves 87.1% F-measure on discourse-new instances in MUC6, which is better than the results in Bean and Riloff's work. Integrating the discourse-new classifier into their coreference resolver, they get an improvement in precision, but Recall - and F-measure - suffer, especially for non-pronominal mentions. By modifying the system so that noun phrases with string-matching and name variant (alias) antecedents are always resolved, they are able to get back some of the recall while not sacrificing too much precision, and they are able to actually get an improvement over the version without discourse-new classification.

Large-scale pattern statistics Uryupina (2003) proposed to use the Web for identifying uniquely specifying definite noun phrases such as those identified by Bean and Riloff (1999), in a more domain-independent fashion. As features for her classifier, she includes the following ratios of web counts:

$$\frac{\#\text{the } Y}{\#Y}, \quad \frac{\#\text{the } Y}{\#\text{a } Y}, \quad \frac{\#\text{the } H}{\#H}, \quad \frac{\#\text{the } H}{\#\text{a } H}$$

(where Y denotes the noun phrase without the determiner, and H denotes just the head noun).

Together with syntactic heuristics and same-head resolution, she achieves 86.9% F-measure on discourse-new instances and 92.0% F-measure on the task of identifying non-unique noun phrases, which would be more similar to Bean and Riloff's original task.

Results for detecting discourse-new instances as an isolated task leave open the question if its inclusion would be an actual benefit to a system. Ng and Cardie's experiments, where a discourse-new classifier had to be overridden by ad-hoc constraints for the system to benefit from it, show that the integration of discourse-new information into a system is not trivial.

Poesio *et al.* (2005) pursue the approach of discourse-new detection further with the integration into GuiTAR, which implements Vieira and Poesio's heuristics-based approach to same-head coreference together with a classifier and achieve an improvement in the overall resolution accuracy on the Gnome corpus.

Newer versions of GuiTAR (Kabadjov, 2007; Steinberger *et al.*, 2007) also include counts $\frac{\#\text{the } A}{\#A}$, $\frac{\#\text{the } A}{\#\text{a } A}$ for the first premodifying adjective. Results on different corpora, including the Vieira-Poesio corpus and Gnome indicate that discourse-new classification with a rich feature set improves the performance of same-head definite description resolution by a small but significant amount; Kabadjov was able to get improvements for a variety of classifiers, but found that SVMs and multilayer perceptrons worked best.

4.3 Integration of Semantic Features

Most approaches to coreference resolution, including (but not limited to) the ones of Soon *et al.* (2001), Ng and Cardie (2002c) and Yang *et al.* (2003), do not attempt to resolve coreference relations between mentions that have different heads, such as *a car*... *the vehicle*, or *France*... *the country*. One reason for this is that the resolution of definite descriptions – even where head equality gives sufficient evidence for the equality of semantic types – is much less precise than the resolution of anaphoric pronouns, or name matching; so improving recall in the resolution of definite descriptions may actually hurt the performance of the system as measured by an evaluation metric. The other reason is that resolving these non-same-head coreferent definite descriptions – cases which we will label as *coreferent bridging* in the rest of the text – necessitate the use of some form of world knowledge, which involves not only the problem of representation of the world knowledge, but also the problem of how world knowledge is used (and, more generally, how approximations to human world knowledge can be used) in discourse processing.

Since we are most interested in semantic features that will allow us a greater precision in the coreference resolution of full noun phrases (as opposed to pronominal anaphora), I will mention only in passing the work of Dagan and Itai (1990) and Kehler *et al.* (2004), who used predicate-argument frequencies to inform pro-

noun resolution and either found that it resulted in a measurable advantage (in the case of Dagan and Itai) or no advantage at all (in the case of Kehler et al.).

To resolve coreferent bridging, one needs information sources that will help identify similarity, subsumption, or general compatibility of mentions, but it is just as important to answer the question how this information is to be integrated in the system. The problem of discourse-new detection (mentioned beforehand), but also the potential noisiness of information sources, have to be accounted for.

As for the information sources themselves, one can speak of 'traditional' approaches, which use a hand-crafted taxonomy – most commonly, wordnets such as Princeton WordNet and gazetteer lists, which contain a large list of names belonging to one semantic category (such as persons, organizations, airports, etc.) Successful rule-based approaches at the MUC-6 and MUC-7 competitions already exploited path search in WordNet (Lin, 1995) or gazetteer lists organized in a shallow taxonomy (Kameyama, 1997), whereas Vieira and Poesio (1997) look for explicit WordNet relations.

To improve the recall that one gets from these more traditional approaches, one can try and harness semi-formal information that is inherent, for example, in the glosses of WordNet entries (Harabagiu *et al.*, 2001), or the category structure of Wikipedia (which is not a taxonomy as such, but is generally organized according to specificity, cf. Ponzetto and Strube, 2006, 2007).

On the noisy end of the scale, we find approaches that extract information from text – either similarity measures that allow to find synonyms and near-synonyms, or subsumption relations (Gasperin and Vieira, 2004; Markert and Nissim, 2005; Garera and Yarowsky, 2006).

It has generally been argued that the increased noisiness of the approaches based on community-generated content or on information learned from raw text is by far outweighed by the larger recall that one obtains from these sources of evidence. This argument is made even more strenuously in work that uses noisy information sources and evaluates it on hand-picked examples where an antecedent with the desired (lexical/conceptual) relation is present (for example, Garera and Yarowsky, 2006), whereas research that uses such features in a complete coreference system (such as Ng, 2007) is usually more cautious.

This opposition between small-scale evaluations being used to emphasize the value of high-recall information sources, and full-scale systems adopting a much more conservative approach is a general trend in the literature reviewed below. However, it is not completely clear whether this is an engineering issue (as both the information sources themselves and the construction of a complete coreference system present a considerable engineering challenge) or whether the evaluation setting chosen by the proponents of 'new' information sources is simply overly optimistic.

Closely related to the question of the recall/precision tradeoff is that of the (logical, conceptual, or lexical) relation targeted by the approaches, and whether the approach is meant to tackle a specific relation (for example, names taken up again by a definite description such as "*Bremen . . . the city*"), or meant as an ingre-

dient for a general problem of plausible coreference.

Finally, there are approaches that do not start from useful generalizations, but instead extract statistics about the behavior of words with a more direct focus on coreference, which becomes increasingly attractive with the growing size of available referentially annotated corpora: Ng (2007) extracts statistics regarding the anaphoricity of noun phrases, and of likely coreference (allowing to exploit the fact that texts from one time period and country typically have the same antecedent for mentions such as "*the president*"). The approach of Ponzetto and Strube (2006) uses features based on semantic role labeling, which also aims at learning regularities from the annotated corpus itself. While such approaches are empirically successful, the exact nature of the relations learned is neither investigated nor discussed, which is regrettable since they could (potentially) uncover different insights or strategies from those realized by the approaches based on existing resources and/or unsupervised learning.

In chapter 6, I will go into greater detail regarding the underlying techniques used by text-based approaches as a prerequisite for a comparative evaluation of such techniques using German newspaper text.

4.3.1 Using WordNet

Poesio, Vieira, and Teufel (1997) investigate both coreferent and non-coreferent bridging relations between definite descriptions and mentions in the previous text. They break up the bridging descriptions into six classes, motivated mostly by processing considerations:

- lexical relations between the heads:
  synonymy/hyperonymy/meronymy (e.g.: *new album*... *the record*)
- instance relations linking definite descriptions to proper names (e.g.: *Bach* ... *the composer*)
- modifiers in compound nouns (e.g.: *discount packages*... *the discounts*)
- event entities introduced by VPs (*Kadane Oil Co. is currently drilling*... *the activity*)
- associative bridging on a discourse topic (*the industry* in a text on oil companies)
- more complex inferential relations, including causal relations

In the 204 bridging definite descriptions from the corpus they analyzed, 19% had a lexical relation between common noun heads and 24% were definite descriptions referring to named entities.

Poesio, Vieira and Teufel then tried to solve the instances involving lexical relations using WordNet, looking for synonyms, hyperonyms and coordinates sisters in the WordNet graph in the case of synonymy and hyperonymy, and direct

meronyms or meronyms of hypernyms in the case of meronymy. They find that an approach combining the relation information in WordNet with distance information (i.e., resolving to the nearest element that has a compatible distance) yields a precision between 36% (for synonyms) and 20% (for coordinate sisters); overall recall (proportion where a relevant relation between anaphor and antecedent could be found in WordNet) was 39%.

For proper names, they assign a semantic type, based on appositive constructions and honorifics or other name parts, such as *Mr., Co., Inc.* as cues. With a mechanism for propagating the type to other entities (e.g. from *Mr. Morishita* to *Morishita*), they can correctly classify 69% of the names in the corpus, and are able to resolve 52% of the definite descriptions referring to named entities.

Harabagiu *et al.* (2001) go beyond synonymy and hyperonymy and consider more general paths in WordNet that they find between anaphor-antecedent pairs found in the training data. To find candidate pairs, they first remove from consideration anaphoric expressions with an antecedent that can be found with knowledge-poor methods, such as string matching, appositions, name variation, or the most salient compatible antecedent.

For the remaining anaphoric definite expressions, they look for anaphor-antecedent pairs that are related by at most five of the following relation types in the WordNet graph:

- SYNONYM, ISA/R-ISA and HAS-PART correspond to synonymy and hyperonymy and meronymy relations.
- GLOSS/DEFINES connect a word in a synset to the word used to define it.
- IN-GLOSS/IN-DEFINITION connects an element of the synset with one of the first words in its definition.
- MORPHO-DERIVATION connects morphologically related words.
- COLLIDE-SENSE connects synsets of the same word (homonyms or polysemous senses).

Harabagiu *et al.* then use the figure of merit outlined below to add some of the paths as rules: they collect all paths where the majority of instantiations found in the training data has a figure of merit greater than the threshold, these paths are then added as rules, and the procedure repeated with the resulting new set of anaphoric expressions and a lower threshold for the figure-of-merit until the overall F-measure drops below a certain point.

The figure of merit for Wordnet paths is defined as the weighted harmonic mean of two factors (where larger numbers mean better confidence):

The first factor (named f_2 in their paper since they also have an additional factor f_1, mentioned below) prefers "stronger" relations: each relation type is assigned a weight $w(rel)$ (ranging from 1.0 for SYNONYM over 0.9 for ISA and GLOSS down to 0.3 for IN-GLOSS), and the weight is averaged over the relation

types occurring in the path, with multiple occurrences of a relation weighted down by a factor corresponding to their number of occurrences ($n_{same}(\mathrm{rel})$):

$$f_2 = \frac{1}{n_{rel}} \sum_{rel \epsilon Path} \frac{w(rel)}{n_{same}(rel)}$$

Additionally, the total number of different relations (n_{rel}) is used to weight down longer paths.

As an example, a path with one HASPART edge (weight 0.7) and two ISA edges (weight 0.9) would receive a weight of $\frac{1}{2} \cdot \left(\frac{0.7}{1} + \frac{0.9}{2}\right) \approx 0.57$, whereas a path with two ISA edges would receive a score of $\frac{1}{1} \cdot \frac{0.9}{2}$.

The second factor, f_3, is determined by considering the search space built when considering at most five combinations of the relations above, starting from either of the nodes. The first factor contains the proportion of nodes on the path joining the two words with a maximal f_2-score ($n_{path}(SS)$), to the total number of nodes on paths joining the two words ($n_{total}(SS)$). The second factor $\log \frac{C}{N}$ (more likely $\log \frac{N}{C}$ or simply $\frac{C}{N}$)[7] depends on the ratio between the intersection of the sets of nodes reachable by anaphor and (candidate) antecedent (with size C) to the total search space, i.e., the union of the reachable node sets (with size N).

$$f_3 = 0.5 + \frac{0.5 n_{path}(SS)}{n_{total}(SS)} \log \frac{C}{N}$$

Harabagiu *et al.* use a weighted harmonic mean of their f_2 and f_3 (with f_3 weighted about 6.7 times higher than f_2), plus an additional term of (about) 1.0 if both NPs have a possessor (either a possessive pronoun or a genitive NP) and these possessors are already coreferent.

Harabagiu *et al.* do not offer a motivation for the exact choice of the many weighting functions employed by them – as a consequent, it is not always clear whether the exact choice is due to an intuition about the problem to be solved or the result of ad-hoc modifications to a function until a satisfactory result is reached. At the same time, their work contains several noteworthy insights:

- Harabagiu *et al.* do not always use the direct antecedent (i.e., closest previous member of the coreference chain, which could require a name to be resolved to a pronoun), but instead allow their system to learn a relation to an antecedent that is further away. (Which is also the case in more rule-based approaches such as the one by Poesio *et al.*, 1997, but not in many machine-learning approaches such as those of Soon *et al.*, 2001 or Ponzetto and Strube, 2006).

[7] Since $\frac{C}{N}$ should be smaller than 1 in most cases, its log would be less than zero, which does not make sense. If one is to take Harabagiu et al.'s claim that the formula is "*inspired by Salton and Buckley's tf-idf weighting*" at face value, dividing the larger N by the smaller C would be a more convincing analogue to document frequency; if one conversely claims that it is better if the search spaces have a considerable overlap, $\log(1 + \frac{C}{N})$ or simply $\frac{C}{N}$ would be a plausible reading.

- They use WordNet to derive a general distance measure including the definitions contained in the glosses, yielding a markedly different information source from Poesio et al.'s earlier approach that is more focused on using the information in WordNet as it is and getting highly precise subsumption and synonymy predictions.

- Harabagiu *et al.* use a global clustering-based model that can make use of more reliable decisions (e.g. for possessive pronouns) to influence other decisions (for the possessed NPs) where the coreference between the possessors provides additional information.

Harabagiu *et al.*'s results (cf. table 4.1) – a MUC F-measure of 81.9% – are below Luo's results for the merge-everything baseline.

4.3.2 Acquisition from Text

The coverage of hand-crafted taxonomies – whether designed with great care or the result of the unstructured collaboration of a multitude of contributors – will always be limited in the sense that adding additional concepts is only possible by manually inserting them.

The alternative to these formally constructed resources are approaches that take a large quantity of text as input and use the (linguistic) contexts in which a word occurs to derive a representation that can be used effectively as a proxy for the meaning of the word in applications such as coreference resolution.

Work in ontology learning, such as learning of taxonomies (Caraballo, 1999; Cederberg and Widdows, 2003; Snow *et al.*, 2006), relation learning (Hearst, 1998; Girju, 2003; Cimiano and Wenderoth, 2005) and ontology population (Thelen and Riloff, 2002; Cimiano and Staab, 2004) aims at a roughly similar goal – automatically constructing artifacts that serve a similar purpose that hand-crafted taxonomies would – and develop techniques that may also be fruitfully used for the above purpose. However, it is not advisable to turn to such approaches and expect the finished results of such work to be usable off the shelf; One reason is that such approaches often trade in too much precision for recall, in addition to the fact that research in ontology learning seldomly uses standardized evaluation procedures.[8] More often than not, ontology learning approaches are evaluated on a small, well-behaved test set, and the question of suitability for larger-scale use is not addressed. A further reason why an intermediate representation would be more relevant to the task at hand than a finished, ontology-like product is pointed out by Markert and Nissim (2005): some relations between mentions, such as *age* being a *risk factor* are only relevant in specific contexts, and would be undesirable in a

[8] Buitelaar *et al.* (2005) put this optimistically as "*... the field is rapidly adapting the rigorous evaluation methods that are central to most machine learning work. Therefore, ontology learning will be impacted by efforts to systematically evaluate and compare approaches on well-defined tasks and with well-defined evaluation measures, thus making it a highly challenging field in which only competitive and demonstrable approaches will survive.*" Note the use of future tense.

general-purpose ontology, which means that an intermediate representation may be more suitable for a coreference system.

Within the general problem of summarizing the (linguistic) contexts in which a word occurs into a representation that can be used as a proxy for word meaning, different techniques can be used to represent the contexts and exploit the associated statistics in ways relevant to coreference resolution:

- The first, constructing a vector space out of the context statistics, can be seen as targeting mostly *similarity* relations; integration with a coreference system can be realized either by imposing similarity thresholds (either absolute or by creating a list of most-similar items), in which case the information can be used as a binary flag, or as a continuous similarity score, in which case it can be used for ranking antecedents.
- The second solution directly targets is-a-relations that would be helpful in detecting logical *subsumption* between two mentions, by using patterns that are strong indicators of these (lexical) relations.
- Finally, the third method directly targets anaphor-antecedent relations by collecting plausible antecedents to a (presumedly) anaphoric definite and uses statistical co-occurrence measures to reduce false positives.

Co-occurrence statistics Poesio, Schulte im Walde, and Brew (1998) investigated the use of second-order co-occurrences of words with a fixed-size context window (see section 6.1 for a broader overview of such techniques), on the same data set as Poesio, Vieira and Teufel (1997). Using the count vector of words co-occurring with a given word and different vector distances, they used the British National Corpus (Burnard, 1995) to learn an association measure of words. They tried out different combinations of window sizes and similarity metrics, as well as a variant with lemmatization and part-of-speech marking. In the best configuration they found, they got 22.2% precision for synonymy/hyperonymy/meronymy cases overall, against 39% for the WordNet-based approach. The more complex inferential relations were the only area where the association measure would outperform the more precise alternative methods.

Distributional similarity Gasperin and Vieira (2004) use a word similarity measure (Gasperin *et al.*, 2001) very similar to the one introduced by Lin (1998a).[9] In contrast to Poesio, Schulte im Walde, and Brew's work, they do not resolve to the semantically closest noun, but instead build lists of globally most similar words (a so-called *distributional thesaurus*), and enable the resolution to antecedents that are in the most-similar list of the anaphoric definite, where the antecedent has the anaphoric definite in its most-similar list, or where the two lists overlap. Working

[9] See section 6.1.1 for a more extensive discussion of Lin's distributional similarity measure.

on Portuguese data, Gasperin and Vieira find that they reach similar levels of resolution accuracy to the results of Poesio, Schulte im Walde and Brew's (1998) with a window-based association metric.

Seeing the results of Poesio et al., and of Gasperin and Vieira, we find several unanswered quesions: one is the relative merits of association measures such as HAL (used by Poesio, Schulte im Walde, and Brew) versus distributional similarity metrics, but also about the relative merits of either ranking by the similarity estimate or using it to create lists of most-similar words and using these as a binary distinction.

Pattern-based approaches Another line of work uses shallow patterns to target specific relations (hyperonymy and, less interesting for our purposes, meronymy). The advantage of these shallow patterns is that they can be made simple enough to be queried using search engines.

Poesio, Ishikawa, Schulte im Walde, and Vieira (2002) use patterns such as "*the NP of NP*" or "*NP's NP*" to extract additional meronymy relations from the BNC that are not found in WordNet or by their vector-based model, and find that the pattern-based model performs considerably better than either WordNet or the vector-based model for resolving cases of associative (meronymy) bridging.

Markert, Nissim, and Modjeska (2003) are concerned with the question of '*other*-anaphora', where a noun phrase such as "*other risk factors for Mr. Cray's company*" is to be linked back to the anchor, the risk factor of "*the designer's age*". To find these relations, which are clearly subsumption relations on a logical/conceptual level, they propose to search for patterns such as "*NP and other NPs*", based on search engine queries using Google. Markert *et al.* then compute a mutual information statistic between the NPs to prefer antecedents with a strong association over frequent antecedents with some spurious matches. They find that for a sample of 120 *other*-anaphora, the web-based technique performs slightly better than an existing approach using WordNet.

In a second experiment, Markert *et al.* investigate the use of Web-based pattern search for associative bridging and find that using raw counts for the World Wide Web performs on a similar level than Poesio *et al.*'s (2002) earlier approach using the BNC.

In a subsequent article, Markert and Nissim (2005) do a more thorough investigation, comparing the use of WordNet and pattern-based search both using the BNC and search engine queries for the World Wide Web for the purpose of resolving coreferent bridging cases. In an experiment where the antecedent for a discourse-old definite noun phrase is to be found, including both same-head cases and coreferent bridging, they achieve F-measure results of 66.4% for string matching with number checking and replacing the string of named entity antecedents with their NE class (person, company, or location), against 69.7% for first looking for a string-match antecedent and then looking for hyperonyms in WordNet, 67.8% using pattern search on the BNC and 71.4% using pattern search on the

World Wide Web, showing web searches to be more effective due to the greater recall that is achieved.

Markert and Nissim (2005) simply use the frequency count of pattern matches to break ties between multiple antecedents with a pattern, or recency to break ties between multiple string-matching antecedents.

Other work emphasizes different methods to integrate different information sources such as pattern search, WordNet information, or salience indicators (including recency and grammatical function): Modjeska *et al.* (2003) use the sample selection and resolution algorithms of Soon *et al.* (2001) with a naïve Bayes classifier to resolve *other*-anaphora; besides features commonly found in coreference systems such as type of noun phrase, grammatical role, syntactic parallelism, distance, semantic class and agreement of semantic class as well as gender agreement, they add one feature indicating the relation found in WordNet, and also two binary features *webfirst* (for the noun phrase with the highest mutual information value from the pattern-based search on the World Wide Web) and *webrest* (for all other possible antecedents where a pattern match could be found). Modjeska *et al.* find that resolution using both WordNet and Web patterns works considerably better than only using the WordNet-based feature.

In a similar fashion, Poesio, Mehta, Maroudas, and Hitzeman (2004a) use a multilayer perceptron with features including simple graph distance in WordNet (indicating the number of nodes between the anaphor and the potential antecedent) and a feature based on the raw count of matches for a search engine query using a meronymy pattern. To express salience, Poesio *et al.* include the sentence distance to the anaphor, but also whether it is in first-mention position, or if any preceding mention of the entity had been in first-mention position.

Using a 1:3 subsampling of positive vs. negative instances (since use all possible antecedents for training and not just the ones nearer than the actual antecedent as in Modjeska's experiment, they need to apply subsampling to reduce the overabundance of negative examples), the best classifier reaches an F-measure of 0.5 on the resolution of bridging examples on the Gnome corpus (which is different from the Wall Street Journal examples used by Poesio *et al.*, 1997 or Poesio *et al.*, 2002).

Daumé III and Marcu (2005) use several classes of features. Besides including WordNet graph distance and WordNet information for preceding/following verbs (in an attempt to let the coreference resolver learn approximate selectional preferences in a supervised way), they also use name-nominal instance lists mined from a large news corpus (Fleischman *et al.*, 2003), as well as similar data mined from a huge (138GB) web corpus (Ravichandran *et al.*, 2005). They also used several large gazetteer lists of countries cities, islands, ports, provinces, states, airport locations and company names, as well as a list of group terms that may be referenced with a plural term.

Contextual co-occurrence as relatedness A third strategy to acquire information that is relevant to the resolution of bridging anaphora is the assumption that the regularities that underly bridging anaphora in general can be uncovered using statistics that find associations among the headword of a (potential) anaphor and plausible antecedents.

Like Poesio et al.'s window-based similarity, but unlike WordNet search or pattern-based approaches, this strategy will discover *relatedness* as the base for a potential anaphoric relation rather than making any specific relation explicit. As a result, the approach has been proposed both for associative and coreferent bridging (and indeed may not discriminate between coreferent and non-coreferent bridges). Unlike the methods based on similarity measures, however, these discourse-based first order associations are directed. They are in principle able to acquire non-symmetrical relations such as meronymy or subsumption.

One implementation based on this intuition is the work by Bunescu (2003), who encodes the intuition into a general web pattern of the type "*X. The Y [verb]*" (where *verb* is an auxiliary or modal verb), assuming that *X* would be a salient possible antecedent for *Y*. Bunescu uses a mutual information statistic, comparing the counts for the pattern instance to those for "*X.*" and "*The Y [verb]*" (where *verb* is an auxiliary or modal verb).

On a test set of associative bridging anaphora sampled from the Brown corpus section of the Penn Treebank, Bunescu's approach reaches a precision of 53% at a recall of 22.7%.

A very similar approach is chosen by Garera and Yarowsky (2006), who investigate the use of bridging-indicating associations to find likely categories for named entities. Using the English Gigaword corpus, they also evaluate a hand-selected sample of coreferent bridging cases and come to the conclusion that, when using the same corpus, their association measure performs better than Hearst-style patterns.

Role co-occurrence and selectional preferences The three general approaches mentioned above learn relations between the head lemmas, independently of the annotated data for coreference resolution. Outside this paradigm of using unsupervised learning to acquire an information source for the head lemma relations that occur in coreferent (and non-coreferent) bridging, several approaches integrate information besides the head word – selectional preferences in the case of Bean and Riloff (2004), relations extracted from an external system in the case of Ji *et al.* (2005) – or directly use the training data to learn relevant distinctions.

Bean and Riloff (2004) use statistics on semantic classes and lemmas with verbs (selectional preferences) and co-occurrence of verb roles (i.e., X was deported – X was freed). They collect lexical expectations based on the syntactic context, as well as expectations about pairs of contexts that might hold coreferent mentions. Bean and Riloff remark that, if a noun phrase X is coreferent with a noun phrase Y in the text, they should be (loosely speaking) substitutable. In their

example

(4.5) a. Fred was killed by a masked man with *a revolver*.
b. The burglar fired *the gun* three times and fled.

it would be acceptable that Fred was *killed with a gun* and that the burglar *fired a revolver*, which is supported by the fact that *revolver* has been extracted by the caseframe for *fire* ⟨NP⟩. Bean and Riloff extract co-occurrences of caseframes and nouns for the lexical expectations, and co-occurrences of caseframes (mediated by named entity coreference) for the caseframe relations. They then use the log-likelihood measure of Dunning (1993) to calculate a χ^2 statistic of the co-occurrences and transform the χ^2 statistic to a confidence value (i.e., the likelihood that a caseframe-and-noun, or caseframe-and-caseframe pair co-occurs more often than by chance).

Bean and Riloff test their approach by including the features into a coreference system that combines multiple preferences in a Dempster-Shafer model[10] (Dempster, 1968; Shafer, 1976), using two test sets containing terrorism-related texts from the MUC-4 evaluation and natural disaster-related texts from the Reuter's news text collection, mentioning that open-domain test corpora such as MUC-6 and MUC-7 would bring the problem of insufficient training set size. Although Bean and Riloff's approach does not depend on annotated training data, it is still likely that it is dependent on a large quantity of unannotated training texts closely matching the genre and domain of the testing texts (the unsupervised training data comprised 1600 unannotated MUC-4 texts and 8245 newswire texts on natural disasters).

In the results of their blind evaluation, they find that although all additional knowledge sources increase the recall of pronoun resolution and most give a small increase in the recall for definite noun phrases, the additional information gives a visible benefit only for pronouns.

Another approach to go past the similarity of heads by using selectional preferences is the one of Castaño *et al.* (2002), who present a semantically informed approach to resolve pronouns and definite noun phrases in medical text. Assigning semantic classes such as *protein* or *amino acid* to noun phrases, they can not only use this information to resolve definite nominals such as *the protein*, but also use selectional preferences in the resolution of pronouns; for example, the agent of *inhibit* would be coerced to one of *amino acid*, *peptide*, *chemical*, *organism function*, or *nucleic acid*.

On a test set including 116 unique antecedent-anaphor pairs in 54 biomedical abstracts from MEDLINE, Castaño *et al.* achieve a precision of 80% and recall of 71% for third-person pronouns and precision of 74% and recall of 75% for

[10] Dempster-Shafer models assign a probability distribution not on single outcomes $\omega \in \Omega$, but on classes of outcomes $A \subseteq \Omega$, with the interpretation that each single outcome ω can be assigned a *minimal probability* (belief) and a *maximal probability* (plausibility) that could occur when additional facts become known. It is also possible to merge two such models by applying Dempster's rule, where a new distribution is formed from the intersections of sets from these distributions in a fashion comparable to Naïve Bayes models.

sortal definite descriptions, against a baseline of P=76%/R=67% for third-person pronouns and P=74%/R=74% for sortal definite descriptions.

Using relation extraction Ji *et al.* (2005) use heuristics to integrate constraints that follow from relations extracted from the text: using an ACE relation tagger, they learn rules that indicates that a pair of mentions 1B and 2B is more or less likely to corefer, knowing the respectively related mentions 1A and 2A with their relations and whether 1A and 2A corefer.

They posit three kinds of rules:

1. If the relation between 1A and 1B is the same as the relation between 2A and 2B, and 1A and 2A don't corefer, then 1B and 2B are less likely to corefer.
   $Same\ Relation \wedge \neg CorefA \Rightarrow CorefBLessLikely$

2. If the relation between 1A and 1B is different from the relation between 2A and 2B and there is coreference between 1A and 2A, then 1B and 2B are less likely to corefer.
   $\neg Same\ Relation \wedge CorefA \Rightarrow CorefBLessLikely$

3. If the relation between 1A and 1B is the same as the relation between 2A and 2B and there is coreference between 1A and 2A, then 1B and 2B are more likely to corefer.
   $Same\ Relation \wedge CorefA \Rightarrow CorefBMoreLikely$

While rule 2 usually has high accuracy independently of the particular relation, as Ji *et al.* put forward, the accuracy of the rules 1 and 3 depends on the particular relation shared among the pairs 1A/B and 2A/B – both on properties of the relation involved and on the reliability of the coreference relations that come into the equation. (For example, the chairman of a company, which has a EMP-ORG/Employ-Executive relation, may be more likely to remain the same chairman across the text than a spokesperson of that company, which is in the EMP-ORG/Employ-Staff relation to it).

Because the rules are differently reliable depending on how they are instantiated, Ji *et al.* look at rule instantiations where only one particular ACE relation is considered. For a rule instantiation to be selected, they require a precision of 70% or more, yielding 58 rule instances. For instances that have a lower precision, they try conjoining additional preconditions such as the absence of temporal modifiers such as "current" and "former", high confidence for the original coreference decisions, substring matching and/or head matching. In this way, they can recover 24 additional reliable rules that consist of one of the weaker rules plus combinations of at most 3 of the additional restrictions.

In Ji et al.'s system, a baseline coreference resolver provides the coreference information that is used in the rules together with the output of the relation tagger, and in a second stage the original information is used together with the constraint

instances to do another pass of the coreference resolution with a model that has been trained on data including the relational features.

They evaluate the system, trained on the ACE 2002 and ACE 2003 training corpora, on the ACE 2004 evaluation data and provide two types of evaluation: the first uses Vilain et al's scoring scheme, but uses perfect mentions, whereas the second uses system mentions, but ignore in the evaluation any mention that is not both in the system and key response. Using these two evaluation methods, they get an improvement in F-measure of about 2% in every case. In the main text of the paper, Ji *et al.* report an improvement in F-measure from 80.1% to 82.4%, largely due to a large gain in recall. These numbers are relatively high due to the fact that Ji *et al.* used a non-standard evaluation setting (discussed as *Disregard spurious links* in section 3.2.5). More realistic results given in a footnote indicate an improvement in F-measure from 62.8% to 64.2% on the newswire section of the ACE corpus, which is still respectable.

Supervised learning of semantic distinctions Learning directly from the training data can make eminent sense where the training data is large enough to learn relevant conventions of the domain: For example, newspaper text usually cites one person for several statements (yielding "Peter said ... He claims... The outspoken CEO denied that ..."), which means that multiple subjects of reporting verbs often corefer. Another example are entities that occur frequently in the text, such as "Bill Clinton" frequently being mentioned as "the President" subsequently in US newspaper texts of the late 1990s.

Ponzetto and Strube (2006) explore features based on semantic role labeling and different distance measures on the Wikipedia category graph and the WordNet hyperonymy graph. In their evaluation, which discards all system-generated markables that are not present in the key, they find that WordNet-based, Wikipedia-based and SRL-based features all bring improvements over their Soon et al.-style baseline. Using wrapper-induced backward feature selection, they are able to get rather drastic improvements on the BNEWS section of the ACE 2003 corpus and visible improvements on the NWIRE section. Because of their nonstandard approach to learning and evaluation (they filter out every markable that is not in the gold-standard before doing coreference resolution proper), it is not clear if these improvements would be realized in a process with completely automatic markable creation.

Ng (2007) includes an ACE-specific semantic class feature that achieves superior results to Soon et al.'s method using WordNet by looking at apposition relations between named entities and common nouns in a large corpus to find better fitting semantic classes than using WordNet alone. In addition, he uses a semantic similarity feature similar to the one introduced by Gasperin *et al.* (indicating if one NP is among the 5 distributionally most-similar items of the other), and two features that are learnt from a subset of the training data that has been set aside:

- a *pattern-based* feature, for which the span in between mentions is encoded

in several patterns with varying degrees of abstraction:

- one that represents NP chunks by the token NP, leaves verb and adverb tokens tokens in the full word form, and replaces any other tokens by their POS tag
- one additionally replacing verb tokens by their POS tag
- one retaining only the NP chunk tokens.

This information, together with agreement properties and mention type of the mentions involved, the accuracy of the patterns is determined using all applicable matches in the development corpus.

The feature value for the pattern-based feature is then the accuracy of the most accurate match for the span between the two noun phrases.

- an *anaphoricity* feature which encodes how often the NP was seen as a discourse-old noun phrase in the corpus, substituting a negative value when the NP has not been seen before.

- a *coreferentiality* feature that encodes the probability that two noun phrases are coreferent, estimated by looking at pairs occurring in the corpus.

Training on the whole ACE-2 corpus, Ng is able to improve from F=0.620 for MUC scoring on the merged test set to F=0.645 for all the features except the pattern-based one.

Yang and Su (2007) present an approach to supervised selection of patterns to be used as features: starting from coreferent pairs from the training data such as "*Bill Clinton*" and "*President*" (or, due to the annotation scheme of the ACE corpora, "*Beijing*" and "*China*"), they extract patterns in Wikipedia as pieces of text that occur between the two mention strings (such as "(Bill Clinton) *is elected* (President)").

Yang and Su propose different methods for using the information from the pattern matches as classification features. As a first step, they filter out patterns based on precision: any pattern that extracts more known-non-coreferent word pairs than known-coreferent word pairs in the training data is discarded.

In a subsequent step, patterns are ranked (either on raw frequency or on the reliability score explained below) and the 100 top-ranking patterns are kept. In the case of the frequency-based approach, a feature is created for each pattern that indicates the frequency of that particular word pair with the pattern on the Wikipedia data. For the other approaches, they calculate a reliability measure for each pattern (determined by summing the pointwise mutual information values between a pair of noun phrases and the pattern, over all coreferent pairs from the ACE training data). The score for a given pattern and a given pair of fillers is then determined as the reliability measure of that pattern multiplied by the positive mutual information between positive mention pairs. Yang and Su propose one

variant where a feature is created for each pattern, and one where the values for each pattern are simply summed to yield a single numeric value.

Yang and Su apply these features in a coreference resolution system similar to the one described by Ng and Cardie (2002c) on the ACE-2 corpus. Using the reliability-based single relatedness feature for proper names (the setting they found to work best) results in an improvement from 64.9% F-measure to 67.1% on the newswire portion, 64.9% to 65.0% on the newspaper portion, and from 62.0% to 62.7% on the broadcast news part. (Yang *et al.* also investigated the use of these patterns on all types, which had a larger detrimental effect on precision).

4.4 The State of German Coreference Resolution

If we look at the development of the English-language coreference community, we notice that a few factors have been especially important for the development of the field:

- The availability of a common corpus suitable both for data-driven development and testing, and a common evaluation metric that correlates well enough with actual annotation quality to be able to compare different approaches.

  This can be seen in the fact that there is very little development in terms of coreference of names and definite descriptions that pre-dates the MUC effort, and replication of these results has essentially not been attempted.

- The availability of suitable off-the-shelf processing components, such as partial or full parsers and named entity recognizers, as well as lexical resources such as gazetteer lists, wordnets, etc.

  The experiments of Kouchnir (2004) for German indicate that using preprocessing techniques that would give reasonable results for English leads to a significant number of mis-identified mentions and/or incorrect morphology. As a result, performance figures are lower by as much as 15% with respect to using the gold standard annotation (Kouchnir uses the PCFG-based chunker of Schmid and Schulte im Walde, 2000 and a maximum entropy-based named entity recognizer to perform automatic preprocessing on the Heidelberg Text Corpus).

For German, researchers usually introduced their own annotated data set (cf. Hartrumpf 2001; Strube *et al.* 2002), which means that replicating or comparing to existing results is notably harder, and until recently meant that coreference resolution was hampered by the the fact that no common data set was publicly available. With the publication of the TüBa-D/Z coreference annotation and its availability under a license that is acceptable for most researchers, as well as its considerable size, there is considerable hope that it will emerge as a standard data set for this task.

The state of the art of preprocessing components such as parsers or named entity recognizers also shows a noticeable gap between English and German: while English NER is essentially seen as a solved task, the approaches that are successful for English do not carry over easily to German, due to the fact that both recognition of NE boundaries (which is more difficult due to capitalized common nouns in German) and the classification of named entities are less accurate (cf. Carreras *et al.* 2003).

While chunkers and full parsers do exist for German (see the overview in section 5.3.3), parsing accuracy is hampered by the more complex interaction between morphology and word order in German. Chunking is also more difficult for German than for English due to prenominal clausal modifiers.

In addition, even simple head or head/premodifier matching techniques that boil down to simple string matching in English necessitate the use of a morphological analyzer in German due to morphological variation on one hand and synthetic compounds on the other.

4.4.1 Pronoun Resolution

There is a some literature on German anaphora resolution which focuses exclusively on pronouns. Since it may nonetheless be interesting for the reader, I will give a brief overview over these approaches.

Schiehlen (2004b) implements a host of constraints and saliency features (ranging from grammatical function parallelism over different variations of surface order to hard binding and agreement constraints) and reports F-measure scores of 78.2% (personal pronouns) and 79.0% (possessive pronouns), respectively.

Kouchnir (2004) presents a system based on boosting of decision tree stubs (using a subset of the features used by Strube *et al.*, 2002, described below). In contrast to other research on pronoun resolution, she reports results both on a corpus with hand-annotated morphological and semantic features and a version from fully automatic processing, where using fully automatic processing results in a performance degradation of about 15%.

Stuckardt (2005) has adapted his pronoun resolver ROSANA (Stuckardt, 2001) to the German language. While the software is freely available for research purposes, no published evaluation results exist.

Hinrichs, Filippova, and Wunsch (2005c,b) present two approaches to German pronoun resolution:

The first is a classifier-based resolver based on memory-based learning. After filtering for number and gender compatibility, they extract features pertaining to the location of the antecedent candidate relative to the anaphor (*cataphoric* and *distance*), parallel grammatical functions (*parallel*), co-argument relation (*clausemate*), the form of the pronoun, and features signaling how often a member of the antecedent candidate's coreference chain occurred with a certain grammatical function label. Among the positively classified antecedent candidates, the closest one is chosen; if none is classified positively, the closest subject is chosen if

there is one within the last three sentences. Using this setting, Hinrichs, Filippova and Wunsch reach an F-measure of 77% for referential personal, possessive and reflexive pronouns using cross-validation.

The second system presented by Hinrichs, Wunsch, and Filippova is a reimplementation of Lappin and Leass' (1994) system, adapted to work with German treebank trees from the TüBa-D/Z corpus. After filtering for number and gender compatibility, as well as for binding restrictions, antecedents are ranked by a salience function similar to Lappin and Leass', but adapted to improve the fit to German newspaper text. The rule-based system, named RAP-G after Lappin and Leass' RAP system, reaches an F-measure of 76.6% on the first release of TüBa-D/Z.

4.4.2 Hartrumpf 2001

Hartrumpf (2001) uses a statistical backoff model to choose between antecedent candidates (both for pronouns and for full noun phrases) that have been identified by manually designed rules. Sentences are processed by a rule-based incremental parser (Helbig and Hartrumpf, 1997), with a chunking parse being put in place where full parsing fails. The rules identifying antecedent candidates rely on extensive knowledge in the semantic lexicon HaGenLex (Hartrumpf *et al.*, 2003) to provide information on group-denoting nouns (in the case of coreference with non-matching number), semantic class (in the case of coreferent bridging) and synonymy information, in addition to sentence distance information.

Antecedent candidates are then ranked using a statistical backoff model that backs off to subsets of the candidates until matching examples are found. Additionally, global semantic compatibility over a coreference set is enforced; for this, a beam search over all possible partitions is carried out, pruning solutions with lower overall scores.

The features for the statistical model are the candidate antecedent position and the licensing coreference rule; with this setup, the system reaches 66% F-measure by 12-fold cross-validation on a corpus containing 502 anaphors.

It is remarkable that Hartrumpf's system implements several features (true ranking, global search) that would only occur much later in English-language learning-based coreference systems such as the one by Luo *et al.* (2004).

4.4.3 Strube/Rapp/Müller 2002

Strube *et al.* (2002) adapted the coreference algorithm of Soon *et al.* (2001) to German data: in addition to features like grammatical function and a coarse semantic classification (both of which were added to the data by hand), they use minimum edit distance between mentions to improve the recall on definite descriptions and names, yielding an improvement in recall from 8.71% to 22.47% in the case of definite noun phrases and from 50.78% to 65.68% in the case of names, yielding a MUC-style result of 65.34% F-measure over all coreference links.

Some elements of Strube et al.'s experiments, notably the pre-annotation of perfect (coarse) semantic class and grammatical function values, as well as perfect markable boundaries, are difficult to reproduce in a fully automatic setting and would lead to degraded performance in the case of fully automatic preprocessing (compare with the results of Kouchnir for pronoun resolution, cited above), but since all noun phrase chunks were marked as markables, a severe distortion as in gold-markable-based evaluation is improbable.

4.4.4 Klenner and Ailloud 2008

Although it is more recent than the main body of work reported in this thesis, the approach of Klenner and Ailloud (2008) is worth mentioning since it is the first work with results for the joint resolution of both pronouns and names or definite noun phrases on the TüBa-D/Z corpus. Klenner and Ailloud make use of a feature set including distance features, grammatical functions and grammatical function parallelism, salience, approximated binding constraints, and head matching (somewhat inspired by Hinrichs *et al.*, 2005b), but instead of using classifier decisions directly, they convert the decisions of the memory-based learner into a confidence value (using the number of positive and negative examples among the k-nearest neighbours), and do a global optimization based on a constraint propagation approach where most-confident pairs are considered first and coreference or non-coreference decisions are propagated to enforce consistency/transitivity of the chain relations; the approach is considerably more efficient than previous approaches based on Integer Linear Programming (Klenner, 2007), but is still able to reap visible improvements (about 0.64% absolute improvement in F-measure in the all-mentions setting) over the situation where classifier decisions are applied sequentially without enforcing consistency.

Klenner and Ailloud include an evaluation both for the setting where all mentions are considered, and for a '*true mentions*' settings where only mentions are considered that are part of a gold-standard coreference chain; even though they use an alignment-based evaluation measure, which Luo *et al.* (2004) and others tout as not having the problems of Vilain et al's MUC measure with respect to evaluation on gold-standard mentions, they report differences of as much as 13% in F-measure between these conditions: In the all-mentions setting, F-measure ranges between 63.98% for using a merge-all approach similar to McCarthy and Lehnert (1995), or 65.09% for using the most confident antecedent, and 65.74% for an improved version. Evaluation results in the "*true*" mentions setting, however, range between 76.89% and 78.74% for the baselines using merge-all and most-confident, respectively to 82.50%.

4.5 Summary: Issues in Coreference Resolution

In this chapter, I have attempted to depict the state of the art in coreference resolution, both in terms of the quality that can be reached in systems evaluated on agreed upon corpora, but also in terms of the necessary resources and useful information that contributes to the quality.

The classifier-based approach of Soon *et al.* (2001) reaches results on the level of good rule-based approaches while using ad-hoc heuristics for sample selection in the training phase and a very limited set of features. Despite this apparent simplicity, Soon *et al.*'s approach hinges on a significant engineering effort in the preprocessing and the actual features (especially the alias feature), and efforts that achieve an incremental improvements on Soon *et al.*'s results (Ng and Cardie, 2002c; Yang *et al.*, 2003) see an increase in complexity that is surprising given the intuition that using machine learning should allow for a *simpler* design. Nevertheless, subsequent approaches can be classified according to the improvements they make over the Soon *et al.* baseline.

One axis of variation along the approaches is the mechanism to combine evidence from different sources into a partition of the mentions (coreference chains). Since this is a structured classification task, this involves a non-trivial mapping from features to the inferred output, and, moreover, exact inference in this clustering task is very certainly intractable, meaning that incremental deterministic or beam-search approaches need to be used.

Successful approaches to incorporate global information into coreference resolution, such as the ones by Hartrumpf (2001); Yang *et al.* (2004); Daumé III and Marcu (2005), still base their features on anaphor-antecedent pairs, but use beam search and coherence constraints on the coreference chains built to filter out nonsensical structures. Approaches that rely solely on clusters without looking at mention pairs seem to either suffer from poor performance, such as the mention-entity model of Luo *et al.* (2004), or use the output of an existing system as input, such as the system of Culotta *et al.* (2007).

The other axis of variation is the use of other information sources beyond the techniques based on string matching that Soon *et al.* use – while string-matching techniques provide a useful first start for the task of coreference resolution, it is clear that a more fine-grained approach is needed both for covering name variation and for covering the various ways in which discourse-old definite descriptions are used.

Even in the case of head-matching antecedents, assessing the compatibility of definite descriptions with an antecedent involves assessing the compatibility of modifiers, as in Vieira and Poesio (2000). In the case of non-same-head antecedent candidates, the problem is much more acute, since a same-head antecedent (i) provides a highly likely candidate and (ii) provides good evidence for the definite description to be anaphoric. In the case of non-same-head candidates, we need to exploit available knowledge sources in a reliable way to assess whether they are compatible or not.

A multitude of approaches, reviewed in section 4.3, have been proposed for the task of identifying antecedents to anaphoric definite descriptions: One such approach uses hand-crafted taxonomies, such as WordNet, either limiting the search to hypernyms, synonyms and near-synonyms as in Poesio *et al.* (1997), or generally searching for paths in WordNet, as done by Harabagiu *et al.* (2001) or Ponzetto and Strube (2006).

Since coverage limits of WordNet might still be an issue (coverage of lexemes in WordNet is quite good, but the recall for relations such as hyperonymy or meronymy is comparably low), researchers have also investigated the potential of approaches to automatically acquire and use vector-based similarity measures. Such approaches include window-based approaches (Poesio *et al.*, 1998) or those based on grammatical relations (Gasperin *et al.*, 2004; Ng, 2007), but also first-order associations in discourse, where statistical methods are used to discover highly associated pairs of anaphors and (frequently co-occurring) antecedents, such as in the approaches of Bunescu (2003) and Garera and Yarowsky (2006). Both general approaches – vector-based similarity measures (at least the window-based approach) and first-order associations in discourse – have been proposed for both associative and coreferent bridging. However, the question of how coreferent and associative relations can be separated is not adequately addressed except in the case of distributional similarity based on grammatical relations. Other approaches, such as that of Poesio *et al.* (2002) for associative bridging or that of Markert and Nissim (2005), use patterns for the explicit identification of meronymy (in the case of Poesio *et al.*) or hyperonymy (in the case of Markert and Nissim).

Beyond resources that look at the head lemmas of mentions to establish the compatibility of an anaphor and an antecedent candidate, approaches such as the ones by Bean and Riloff (2004) and Ji *et al.* (2005) take into account information besides the mention's heads by either modeling selectional preferences, modeling relations between contexts (in the sense of grammatical role plus head verb) that may be highly indicative in a script-like fashion, or model the interaction of domain-specific relations and coreference.

A noteworthy aspect of the research reviewed in this chapter is that most work using either advanced decoding/encoding techniques or semantic features uses non-standard evaluation methods, such resolving only discourse-old mentions (what I called the *Perfect Anaphoricity Information* setting in section 3) in the case of Harabagiu *et al.* (2001); Markert and Nissim (2005); Garera and Yarowsky (2006), evaluating only on key mentions in the case of Luo *et al.* (2004); Ji *et al.* (2005) (corresponding to the *No Singletons* and *Disregard Spurious Links*), or on the intersection of system-generated and key mentions in the case of Ponzetto and Strube (2006) (which is a slight variation on the *No Singletons* setting). In contrast, work that uses a realistic setting, such as that of Ng (2007) or Uryupina (2006), uses these information sources in a more cautious manner and achieves improvements that are visible but much less spectacular than those claimed in the less realistic settings.

To explain this disparity, it is useful to consider the review of evaluation measures and evaluation settings in section 3: Beyond the numerical problems with the MUC evaluation scheme detailed there, it should be clear that in the case of the *No Singletons* and similar settings, fewer competing antecedent candidates and a much smaller impact of (necessarily imperfect) discourse-new identification may add up to overemphasize the role of high-recall but possibly noisier information sources.

Chapter 5

Resources for German

In this chapter, I will outline existing resources (both annotated corpora and tools for automatic processing of text) that are either needed for coreference resolution proper, or that are necessary for the processing that serves to extract semantic information from large quantities of text in order to create the corpus-based similarity and association measures that are described in the next chapter.

For **mention identification**, it is necessary to identify the mentions (phrases that introduce a referential index) together with their extent and further information. On the token level, this includes the head lemma (as nouns can display morphological variation) and morphology (as number agreement, and in certain cases agreement on grammatical gender, can be assumed in order to whittle down the set of potential antecedents).

In German, the grammatical gender of inanimate objects does not have to match between an anaphoric definite description and its antecedent, as illustrated in example (5.1-b). For animate entities, however, grammatical gender almost always coincides with natural gender, so that gender agreement for animate entities is a useful constraint for the coreference resolution of definite descriptions, as demonstrated in example (5.1-a). To make this distinction, a coreference system needs a mechanism for *semantic classification* that reproduces at least the distinction between animate and inanimate entities.

(5.1) a. [$_1$ die Krankenschwester] ... [$_{*1}$ der Mann]
[$_1$ *the[fem] nurse*] ... [$_{*1}$ *the[masc] man*]
b. [$_2$ der Stuhl] ... [$_2$ die Sitzgelegenheit]
[$_2$ *the[masc] chair*] ... [$_2$ *the[fem] seating*]

A different set of resources is necessary for features that inform the resolution of coreferent bridging antecedents. The gathering of context statistics from large corpora to create **distributional semantics** models makes two things necessary: on one hand, a relatively large text corpus, ideally with good tokenization and not too many spelling errors (which is often the case with newspaper text, but can be an issue with text collected from the World Wide Web); to extract context statistics from these texts, it is necessary to do further processing, minimally identifying

noun phrases, but also including informative grammatical or semantic relations.

5.1 Lexical information

To get from the surface form of the head noun to the canonical form (lemma), and to retrieve morphological as well as possibly information, it is necessary to use lexical resources such as morphological analyzers.

A **morphological analyzer** yields possible lemmas and morphological properties (such as case, number, gender) for a word. A **semantic lexicon** can be looked up by lemma (i.e., presupposes some kind of morphological analysis), and contains usable semantic information of some form. Finally, most named entities are not contained in semantic lexicons, which is why one usually complements the lexicon with **gazetteers**, rather large lists of names that belong to one semantic category.

5.1.1 Lemmatizers and Morphological analyzers

For morphologically rich languages such as German, the presence of a morphological lexicon that maps the full form back to an entry with lemma and morphological features (including case, number and gender for nouns, or person, number and tense for verbs) is relatively important. Due to lack of interest from the computational linguistics community, however, most of the high-quality components for this task are still proprietary and it is only recently that there are attempts to create freely available, high-quality morphological analyzers.

One resource for morphological analysis is the lexicon in the German grammar included with the the open source WCDG parser (Foth and Menzel, 2006), which uses the paradigmatic classes of Nakov *et al.* (2002) for nouns and hand-coded paradigms for irregular verbs. The WCDG parser handles compound nouns by using a longest-match heuristic, which improves the coverage of the approach but means that coverage of these items is not perfect (moreover, unknown words with common suffixes such as -ion often get spurious analyses such as *(Ablat)Ion* for *Ablation*).

Whereas the WCDG parser uses a full-form lexicon that is generated out of the lemma/paradigm lexicon by a perl script, it is possible to (partially or completely) encode the relevant transformations by finite-state rules.

SMOR (Schmid *et al.*, 2004), which is based on the SFST finite state toolkit, and its predecessor, DMOR (Schiller and Stöckert, 1995), are based on the IMSLex lexicon (which includes paradigm classes and other lexical information).

As SMOR is a derivational morphology, its analyses do not always contain lemmas, but derivation operators which may not correspond to a unique surface form, for example `an<VPART>bieten<V>er<SUFF><+NN>` for *Anbieter*. Therefore lemmatization is only possible by reconstructing the base form from the anal-

ysis by re-generation, possibly using heuristics for the resolution of ambiguities.[1]

While the finite-state toolkit and the implementation of noun and verb paradigms (SMOR in the narrow sense) is open source, IMSLex is only available on request. In contrast to IMSLex, the Morphisto lexicon (Zielinski and Simon, 2008), which is also compatible with SMOR, contains a smaller set of base forms (about 18 000 lemma entries) and is available under a more liberal Creative Commons licence (non-commercial share-alike).

In addition, several companies have developed proprietary tools for lemmatization and morphological analysis:

The Canoo lemmatizer and morphological analyzer is a commercially available tool from the Basel-based Canoo AG which contains about 310 000 full-form entries for German.

The Finnish company Lingsoft Oy offers another commercial tool, the GerTwoL analyzer (Haapalainen and Majorin, 1995), which is based on a derivational finite-state morphology and the Collins German Dictionary, supplemented with additional entries to yield a lexicon with about 85 000 base forms.

5.1.2 Semantic Lexicons

A semantic lexicon contains semantic distinctions (such as, animacy, or humanness), but in the case of wordnets can also contain information that makes it possible to assess more-specific or less-specific relations between word meanings.

The **WCDG parser lexicon** contains a *sort* attribute for nouns that includes common syntactically relevant classes such as numbers, time units, weekdays, months, measures, currencies, but also *beruf* (profession), *human*, and *company*. The degree of coverage of these classifications is certainly not clear (not all nouns and names are classified in this way), but given the simplicity of the classification and the relative liberal license of the WCDG parser, it can provide a useful starting point. (In the TUECOREF system, the WCDG lexicon is only consulted for names, as GermaNet offers larger coverage).

GermaNet (Kunze and Lemnitzer, 2002) is a German wordnet modeled after the English Princeton WordNet (Miller and Fellbaum, 1991). Wordnets are organized around the idea of a *synset*, which models a single concept in the hierarchy pertaining to one part of speech (nouns, verbs, adjectives, adverbs). A synset has one or more *lexical units* which are one word sense that can be used to express that synset.

The meaning of synsets is expressed (in the English WordNet) by a *gloss* which explains the meaning in natural language, and in *semantic relations* between synsets, but also in *lexical relations* which hold between lexical units.

[1] The most important case where the re-generated forms are ambiguous can be found in compounds, as *Kindmutter*, *Kindsmutter* and *Kindermutter* would be mapped to the same analysis by SMOR.

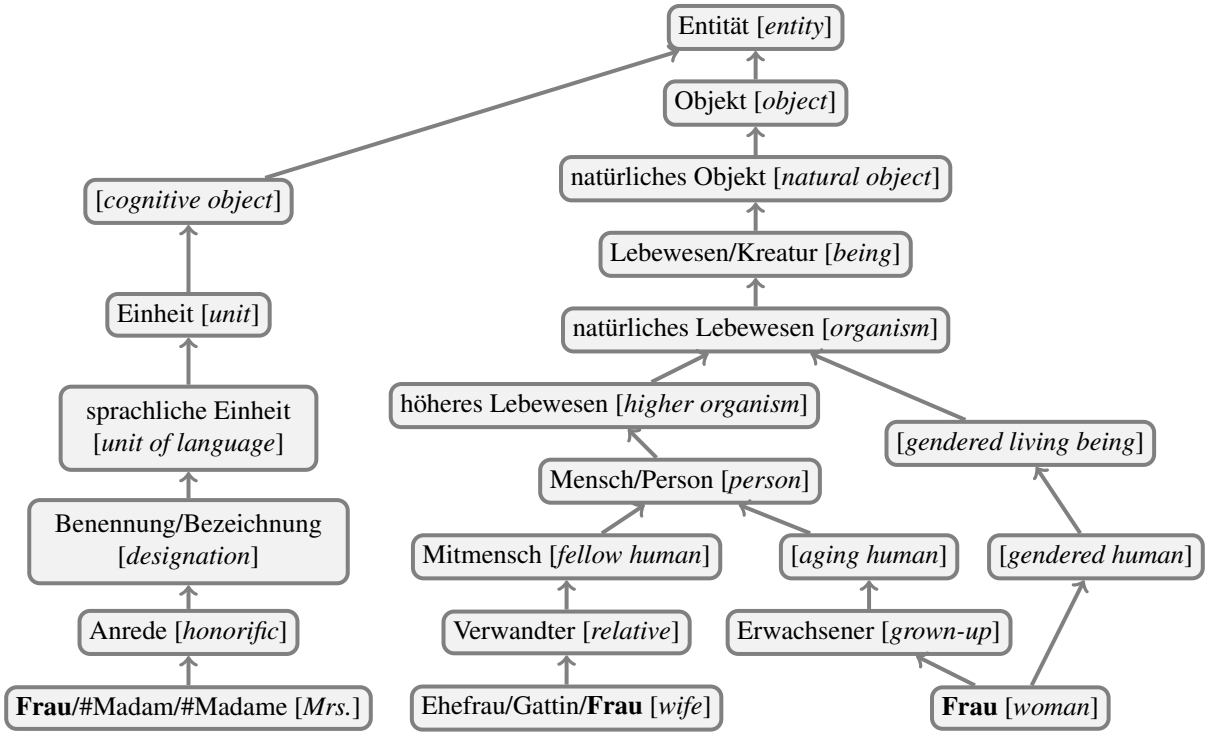


Figure 5.1: GermaNet: senses of *Frau* (wife/woman) with hyperonyms

The usefulness of glosses differs between wordnets: while in the English WordNet, glosses have been used to create to enrich the wordnet relations by parsing the glosses and assigning word senses to the words (Harabagiu *et al.*, 1999), GermaNet contains very few glosses.

Semantic relations in wordnets usually include taxonomic relations such as hyperonymy and hyponymy (less-specific and more-specific relations for nouns) or troponymy (less-specific relation for verbs), but also non-taxonomic relations such as meronymy (part-of relations for nouns) or presupposition relations between verbs such as *succeed* and *try*, which however are annotated much more sparsely than the taxonomic relations. Wordnets also contain so-called *lexical relations* which do not hold between synsets, but between lexical units; the most salient example for a lexical relation is antonymy, where *erheblich* (significant) is an antonym of *unerheblich* (insignificant), but *geringfügig* (immaterial, insignificant), while being a member of the same synset than *unerheblich*, is not an antonym of *erheblich*.

Figure 5.1 shows an extract from GermaNet with the taxonomy represented in the hyperonymy relations between synsets, including different lexical units for a synset. So-called "artificial" lexical units, terms such as '*alternder Mensch*' (aging human) that are needed to explain a synset that does not correspond to a lexicalized concept, have been omitted since their meaning is apparent from the translation. Note that, although relatively uncommon, it is possible for synsets to have multiple hyperonyms, meaning that the taxonomy is (in general) a directed acyclic graph rather than a proper tree.

To assess the semantic similarity of two words, one generally needs to find a path of (taxonomic) relations between the synsets corresponding to one word and the other; the WordNet::Similarity package (Pedersen *et al.*, 2004) retrieves the hyperonym tree for each word, limiting paths to those that go up from one synset to a common ancestor and then down again (meaning that common ancestors, but not common descendants, allow the synsets to have a path between each other). Alternatively, it is possible to use generic graph search methods such as Dijkstra's algorithm (Dijkstra, 1959) – depending on how access to the wordnet graph has been implemented, however, the expansion of aditional nodes can carry a significant cost in terms of computation time and space requirements.

In GermaNet, an additional complication is given by the fact that IT terminology such as "Maus" (mouse), "Monitor" (screen) etc. is marked by linking it with a hyperonymy relation to a synset "*Computerterm*", making it necessary to exclude this synset in the search.

The proprietary **HaGenLex** (used, e.g., by Hartrumpf, 2001) is a high-quality ontology-and-lexicon built according to the principles of the MultiNet formalism of Helbig (2006). It is only used internally by the research group in Hagen and not available to the general public.

Finally, two collaboratively created resources rely on volunteer efforts of the general public for their growth: **OpenThesaurus**, a community effort to build a wordnet based on synonym sets derived from translation and reverse translation of

terms through an open-source bilingual dictionary and subsequent modifications by users. The quality of OpenThesaurus is slowly improving but not in a usable state yet.

Another community project, **Wiktionary**[2], is more oriented towards the structure of a classical dictionary than OpenThesaurus, and with structural consistency enforced through volunteers engaging in clean-up activity rather than restraining entries to a formal model. Zesch *et al.* (2008) show that using the 70,000 entries in Wiktionary as an association net (i.e., not giving any formal meaning to different categories of links, but instead interpreting them as an undirected graph, or using the glosses to build a contextual representation) results in a very usable resource for semantic relatedness.

5.1.3 Gazetteers

For the accurate identification of named entities, it is common to employ gazetteers, large collections of names that belong to one semantic type and allow a more accurate identification of persons, locations, and other entities with names that are not easily recognizable.

For person names, the database of first names and surnames collected by the U.S. Census Bureau[3] is a resource that is frequently used in English. For German, Biemann (2002) has used a bootstrapping approach to acquire a list of first names, surnames, and complete person names that has very good coverage.

The German National Library (Deutsche Nationalbibliothek) maintains a list of person and organization names in a standardized format (Hengel and Pfeifer, 2005). The databases, *Personennamendatei* (file of person names, PND; contains more than one million of person records with a unique identifier and profession) and *Gemeinsame Körperschaftsdatei* (file of organization names, GKD; contains 915,000 records for organization) are available on DVD from the German National Library.

Lists of country and region names are similarly simple to obtain: the United Nations Code for Trade and Transport Locations (UN-LOCODE)[4] assigns a three-letter code to populated places in a country (for example, DE-HAM to Hamburg, Germany) and contains entries for most of the cities that are mentioned in a text. For the processing of German, the UN-LOCODE and other databases with native names only sometimes create problems as German often has germanicized names (such as 'Rom' for Rome, which is called 'Roma' natively). The world-gazetteer database[5] contains population counts, geographic locations, as well as localized names.[6]

[2] http://www.wiktionary.org

[3] `http://www.census.gov/genealogy/names/`

[4] `http://www.unece.org/cefact/codesfortrade/codes_index.htm`

[5] `http://world-gazetteer.com`

[6] The most obvious use for population counts can be found in toponym resolution, where it makes sense to prefer a place with a large population over one with a small population if both are possible

Leidner (2007) recommends larger gazetteers such as those published by the U.S. Board on Geographic Names[7] or the U.S. National Geospatial-Intelligence Agency[8] for the purposes of toponym resolution. However, increasing the size of the gazetteer usually leads to an increasing number of false positives (which would be undesirable unless our goal is toponym classification and resolution with extremely high coverage).

Other resources containing further information on named entities and can be exploited: besides Wikipedia, which contains a large amount of information in a semi-structured format, the Freebase database[9] contains information from Wikipedia (partially corrected and completed by Freebase contributors) in a more principled tabular form (although all in English language).

5.2 Corpora

Corpora – in the loose sense of the word – are collections of texts that have been written for communicative purposes (i.e., text that has been written to convey meaning to a potential or co-present reader/listener rather than sentences created to illustrate or test linguistic theories or their implementation), which can carry additional annotations of various kinds to make underlying linguistic distinctions (more) explicit.[10]

Two kinds of corpora are frequently used in computational linguistics: The first kind consists in smaller, hand-annotated corpora, usually with a size up to one million words, where the annotation of the text is the limiting factor and also the most important aspect. The second kind consists of text collections several orders of magnitude larger, where any kind of manual annotation would be prohibitively expensive. Through automatic annotation, however, it is possible to get many (although not all) of the benefits that hand-annotated would have, and at the same time benefit from the greater size to avoid sparse data effects that would occur with a smaller hand-annotated corpus.

The following sections are not meant to give a representative overview on corpora in general, as it is focused on the resources that are potentially suited to train components in coreference resolution; for a more general overview, the interested reader is directed to the book of Lemnitzer and Zinsmeister (2006).

---

referents for a place name. In semantic classification, the use of population counts is useful to avoid spurious ambiguities between semantic classes that would occur due to small settlements named after persons or other entities.

[7] `http://geonames.usgs.gov/domestic/download_data.htm`

[8] `http://earth-info.nga.mil/gns/html/cntry_files.html`

[9] `http://www.freebase.org`

[10] A narrower view of the term *corpora*, put forward by McEnery and Wilson (1996), requires some care in their collection and the presence of a mixture of genres to ensure representativity. Computational linguistics usually does not subscribe to the same kind of methodological empiricism that corpus linguists see as the foundation to their work. Computational linguists are mostly looking for size and informative annotations in a corpus, with genre diversity, or even any kind of argument for representativity, taking a back seat.

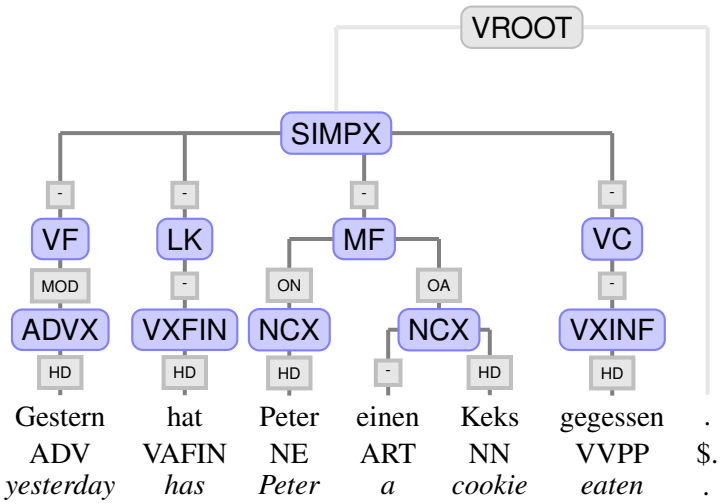


Figure 5.2: Constituent Representations: TÜBA-D/Z

5.2.1 Manually annotated Corpora

Most German corpora that are interesting from a computational linguistics perspective include syntactic annotation, since researchers tend to add new annotation to an existing treebank rather than collecting new text for this purpose. This has the immediate benefits that no additional discussions with rightsholders for the texts are necessary, and that it becomes possible to look for interesting relationships between syntactic and other annotations. Notable exceptions to this rule are the corpus used for the CoNLL 2003 shared task on named entity recognition (Tjong Kim Sang and De Meulder, 2003), which was annotated at the University of Antwerpen and suffers from an overoptimistic application of English named entity guidelines to German (see Rössler, 2006 for a discussion of inconsistencies in the CoNLL 2003 corpus), and the DFKI MuchMore corpus, which contains part-of-speech, chunk, morphology, semantic class, relation, and word sense annotation on a German-English corpus of biomedical abstracts (see Buitelaar *et al.*, 2003; it is unclear how much of the one million word corpus has been manually corrected as opposed to containing purely automatic annotation).

The Tübingen Treebank of Written German (**TüBa-D/Z**; Telljohann *et al.*, 2003) is a corpus consisting of newspaper text from the newspaper *die tageszeitung* (taz), containing words, spelling corrections (separate from the words as they occur in the actual text), part-of-speech tags, morphology (e.g., case/number/gender for nouns), syntactic annotations including phrase structure and grammatical functions as well as secondary edges for certain non-projective attachments (see below for a comparison of the treatment of nonprojectivity in TüBa-D/Z and other annotation schemes), and coreference annotation including separate category labels such

	TüBa-D/Z	NeGra	WCDG
verb arguments			
subject	ON	SB	SUBJ
acc. object	OA	OA	OBJA
dat. object	OD	DA	OBJD
benefactor	OD	DA	ETH
gen. object	OG	OG	OBJG
noun modifiers			
genitive	-	GL/GR	GMOD
postmod. PP	-	MO/OP/PG	PP
apposition	APP[a]	APP	APP

[a]: appositions are handled symmetrically in TüBa-D/Z, whereas they are treated as post-modifiers in NeGra or WCDG.

Table 5.1: Grammatical Functions in NeGra vs. TüBa-D/Z vs. WCDG

as *anaphoric* for pronominal anaphora, *cataphoric* for cataphoric occurrences of pronouns, and *coreferential* for names and other non-pronouns (Naumann, 2006). Additionally, several kinds of non-identity anaphora are marked including reference to groups from entities that werde introduced independently from each other (*split antecedent*), reference to elements of a group introduced earlier (*part of*), and marking of non-referential uses of *es* that are not marked in the syntax.

Release 4 of the TüBa-D/Z covers about 635,000 words of running text in 36,000 sentences, taken from 1,700 newspaper articles, with syntactic annotations containing phrase structure and grammatical functions, containing additional nodes for topological fields (a descriptive scheme to account for the internal structure of clauses; see fig. 5.2 and the discussion on page 131).

The **NeGra** treebank (Skut *et al.*, 1997) is another German treebank and uses dependency grammar along the lines of Hudson (1984) as the foundation of a theory-neutral constituent structure treebank (see figure 5.3 for an example of the NeGra/TiGer annotation scheme). Unlike the TüBa-D/Z, The NeGra treebank does not introduce nonterminal nodes for the phrasal projection of each pre-terminal. Instead, it uses projection nodes for some preterminals to attach dependents to, and adds additional nodes for a symmetric analysis of coordination (i.e., structures of the form $[X\ [X\ w1], [X\ w2]$ and $[X\ w3]]$).[11]

While the NeGra treebank is one of the oldest and also one of the smallest treebanks, with 355,000 tokens in 20,500 sentences, literature in parsing (Dubey and Keller, 2003; Schiehlen, 2004a; Dubey, 2005; Petrov *et al.*, 2006; Petrov and Klein,

[11] A non-symmetric analysis of coordination has been chosen in the TUT-penn conversion of the Torino University Treebank (Bosco and Lombardo, 2006). For a more concrete example, consider the symmetrical analysis [NP [NP noodles], [NP rice] and [NP potatoes]], where the TUT-penn scheme would avoid introducing an additional node for the complete coordination and prefer a chain of conjuncts as in [NP noodles, [NP rice and [NP potatoes]]].

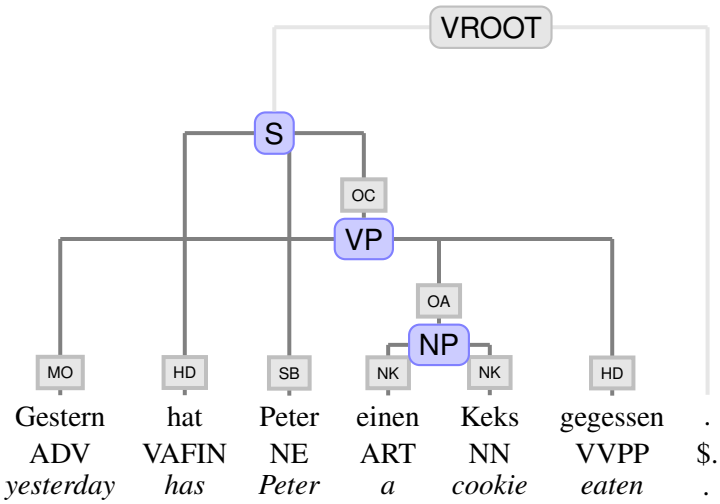


Figure 5.3: Constituent Representations: NEGRA/TIGER treebanks

2008a) obstinately uses it to report results, preferring comparability to prior results over the larger size and richer annotation (especially concerning morphology) of newer treebanks for German.

The **Tiger** treebank (Brants *et al.*, 2002) uses the annotation scheme of NeGra for its syntactic annotation layer, but in addition contains morphological annotation (which has been done semi-automatically and sometimes contains underspecified or overspecific entries), and lemmas. Additionally, current efforts (Burchardt *et al.*, 2006) aim towards the creation of corpus annotation and a frame inventory inspired by the theoretical underpinnings of FrameNet (Baker *et al.*, 1998). Annotation is done by predicate lemma, which mean that all occurrences of a particular verb in the corpus are annotated, but that the released corpus does not contain a complete frame-semantic representation for all sentences.

Differences between syntactic annotation in TüBa-D/Z and NeGra/Tiger The TüBa-D/Z treebank uses a different annotation scheme than the NeGra and Tiger treebanks, which can result in different behaviors of treebank-based parsing strategies (using treebank grammars or history-based parsers), but also means that evaluation figures for parsers trained on TüBa-D/Z on one hand and on NeGra/Tiger on the other hand are more difficult to compare: while both TüBa-D/Z and NeGra/Tiger encode grammatical functions in edge labels (see table 5.1), TüBa-D/Z has a deeper structure than the flat NeGra trees, as NeGra removes unary projections (note that in figure 5.3, the preterminal *Peter/NE* is governed directly by the sentence node), whereas TüBa-D/Z keeps the unary projections (compare with figure 5.2, where the preterminal is governed by an NCX node). Additionally, NeGra has completely flat analyses of noun phrases – yielding a

single NP node for

(5.2) [$_{NP}$ the man [$_{PP}$ with the telescope] [$_{PP}$ in the forest]]

whereas TüBa-D/Z uses an adjunction-like scheme for postmodifiers, which yields multiple NP projections, as in

(5.3) [$_{NX}$ [$_{NX}$ [$_{NCX}$ the man] [$_{PX}$ with the telescope]] [$_{PX}$ in the forest]]

A further difference is to be found at the clausal level: generally, NeGra uses **verb phrases** for further structuring of sentences with auxiliaries (sentences with only a main verb are represented as one S phrase that is completely flat). Due to German free word order, main verb arguments can occur in the pre-field, which means that the VP is no longer contiguous (cf. figure 5.3; a generally accepted model for generative syntax of German posits that the contents of pre-field and the left sentence bracket are moved to specifier and head positions of the complementizer phrase). The NeGra scheme embraces this by completely giving up the notion of projectivity. In consequence, postposed relative clauses and prepositional phrases can simply be attached nonprojectively.

The nonprojectivity of the NeGra annotation scheme means that NeGra and Tiger in their original form are unsuited for PCFG parsing, and have to be made projective by raising offending phrase children to higher nodes. This means that a phrase-structure parser, even if it does recover the edge labels somehow, does not recover the original structure. The non-sufficiency of these projectivized trees is widely ignored, and only few researchers (e.g., Levy and Manning, 2004; Hall and Nivre, 2008) make an attempt to reconstruct the original structure in parsing.

The TüBa-D/Z treebank solves the problem of nonprojectivity using a different approach: instead of using VPs to structure sentence content, it uses the descriptively very successful model of **topological fields** (Erdmann, 1886; Drach, 1937; Höhle, 1986). In the topological fields model, a left and right sentence bracket surround a *middle field* (MF), which contains all the non-extracted VP and IP arguments as well as a *pre-field* (VF), which contains the contents that is extracted to SpecCP in V2 clauses:

(5.4) a. [$_{VF}$ Gestern] [$_{LK}$ hat] [$_{MF}$ Peter einen Keks] [$_{VC}$ gegessen].
[$_{VF}$ *yesterday*] [$_{LK}$ *has*] [$_{MF}$ *Peter a cookie*] [$_{VC}$ *eaten*].
yesterday, Peter ate a cookie.

b. ...dass [$_{MF}$ Peter gestern einen Keks] [$_{VC}$ gegessen hat].
...*that* [$_{MF}$ *Peter yesterday a cookie*] [$_{VC}$ *eaten has*].
...that Peter ate a cookie yesterday.

c. ...den [$_{MF}$ Peter gestern] [$_{VC}$ gegessen hat].
...*which* [$_{MF}$ *Peter yesterday*] [$_{VC}$ *eaten has*].
...which Peter ate yesterday.

This approach solves all the parsing problems due to nonprojectivity in topicalization and scrambling; for the remaining cases – essentially postposed relative

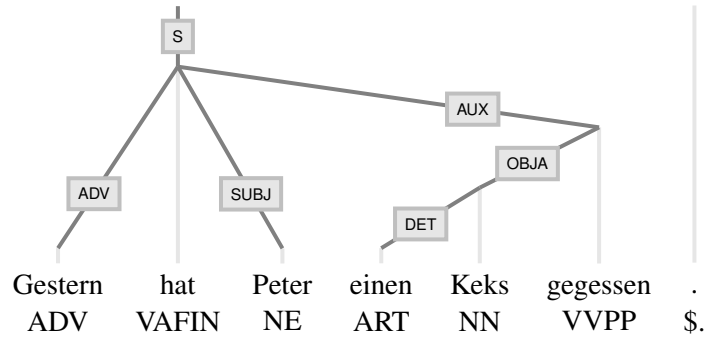


Figure 5.4: Dependency representation: WCDG

clauses, prepositional phrases, or parts of coordinated structures, additional information is annotated in the edge labels or by means of secondary edges. For example the nonprojectively attached relative clause in (5.5) is marked as a modifier of the accusative object:

(5.5) Ich habe [$_{\text{NX.OA}}$ den Staubsauger] gekauft, [$_{\text{R-SIMPX.OA-MOD}}$ den Du
I have the vacuum cleaner bought, that you
Dir gewünscht hast].
yourself wished have.
'I bought the vacuum cleaner you wanted'.

In phrases such as (5.6), where the modifiee of the relative phrase is not in an argument position, a secondary edge would link the NP "dem Kuchen" to its relative clause.

(5.6) Ich habe [$_{\text{PX.V-MOD}}$ von [$_{\text{NX}}$ dem Kuchen]] gegessen, [$_{\text{R-SIMPX.MOD}}$ der
I have of the cake eaten, that
gestern übriggeblieben ist].
yesterday left over is.
'I ate from the cake that was left over yesterday'.

In summary, both treebanks encode similar syntactic information, but in a markedly dissimilar way. The most noticeable effect is that many of the design choices in NeGra (flat analysis of noun phrases and sentences, removal of unary projections) lead to overfitting and sparse data problems, whereas the choices taken in TüBa-D/Z (adjunction-style postmodification, topological fields) reduce sparse data problems, but may lead to underfitting in a PCFG grammar.

Other annotated corpora The Tübingen Treebank of spoken German (TüBa-D/S or Verbmobil treebank; Stegmann *et al.*, 2000) is a collection of utterances from appointment scheduling dialogues, which has been annotated with a scheme similar to that of the TüBa-D/Z (the TüBa-D/S annotation scheme is effectively the ancestor of the TüBa-D/Z's). The TüBa-D/S comprises 18 000 sentences where false starts and disfluencies are handled by annotating partial syntactic structures.

The CDG Corpus Collection (CCC) is a collection of syntactically annotated text including newswire text (Heise newsticker, 100 000 sentences), novels (two fantasy novels, 9 000 sentences), law text (the German constitution, 1 100 sentences), and a sample of Verbmobil dialogues that has been cleaned of false starts and disfluencies (1 300 sentences). The larger part of the newswire text has been annotated in a semi-automatic fashion by correcting parses of the WCDG parser. The annotation scheme used for the CDG Corpus Collection is the same that is also used in the German WCDG grammar and is described by Foth (2003).

The Potsdam Commentary Corpus (PCC) is a small corpus containing syntactic, referential, and discourse structure annotations (Stede, 2004).

5.2.2 Corpora with automatic Annotation

A second group of corpora is *not* manually annotated, but is the result of automatic processing of large quantities of existing texts such as newswire or newspaper archives, parliamentary proceedings, collaboratively created content such as Wikipedia, or an indiscriminate crawl of the World Wide Web.

At such a scale beyond several million words, manual processing is out of the question, and even using automatic processing tools, time consumption can be an issue for the larger corpora. Despite the fact that most such corpora are distributed as raw text or with only the most basic annotation, additional (automatic) annotations can always be added if the need arises and appropriate tools are available.

The Tübingen Partially Parsed Corpus of Written German (**TüPP-D/Z**) is a collection of articles from the newspaper *die tageszeitung* (taz), ranging over all issues from September 1986 through May 1999, comprising about 200 million words. The TüPP-D/Z corpus has been part-of-speech tagged with an ensemble-based part-of-speech tagger, morphologically analyzed using DMOR (Schiller and Stöckert, 1995, see above) and contains phrase and clause chunking as provided by the rule-based chunker of Müller and Ule (2002).

The German web corpus **DE-WaC** (Baroni and Kilgariff, 2006) is a sample of 1.7 billion words of text from the World Wide Web, which has been run through automatic tools for boilerplate removal, spam detection, and deduplication. Linguistic processing for DE-WaC including part-of-speech tagging lemmatization was done using the TreeTagger (Schmid, 1995, see section 5.3.1).

The European Corpus Initiative Multilingual Corpus (**ECI**) contains texts from the German newspaper *Frankfurter Rundschau* from July 1992 to March 1993, totalling about 30 million words.

The Institut für Deutsche Sprache (IDS) in Mannheim curates the **COSMAS-II** (or DeReKo) corpus (Kupietz and Keibel, 2009), containing about 3 billion words (composed for the most part of newspaper text from different newspapers), of which about 2.4 billion words are publically accessible over a web-based interface. While the lack of linguistic processing on one hand and the non-availability of the textual data on the other hand makes this text collection unsuitable for com-

putational linguistics purposes, it is still very popular among proponents of corpus-based linguistics.

Other corpora such as the Huge German Corpus (a corpus of texts from the newspaper *Frankfurter Rundschau* at the University of Stuttgart, comprising 200 million words) and the Wahrig corpus (consisting of 1 billion words from different sources, including newspapers and magazines, which has been collected by the Wahrig publishing company, Saarland University, and CLT Sprachtechnologie GmbH, cf. Pinkal, 2002), are only used internally, but are occasionally referenced in publications from these sites (e.g., Schulte im Walde, 2006).

Finally, a number of documents from parliament proceedings and legal texts from multilingual entities such as the European Union are, thanks to efforts of the statistical machine translation community, available to researchers under relatively liberal terms. One of these is the **EuroParl** corpus (Koehn, 2005), consisting of proceedings from the European Parliament (including text that was spoken in other languages and subsequently translated), which contains 37 million words (or 1.5 million sentences). Another parallel corpus, the JRC/**Acquis Communautaire** corpus (Steinberger *et al.*, 2006), contains a collection of law texts (the legal norms applicable in EU member states), and comprises 32 million words.

5.3 Processing tools

Linguistic processing using automated tools – be it part-of-speech tagging, lemmatization (covered in an earlier subsection), chunking, or full parsing, can help the precision of searches targeted at certain linguistic phenomena, but also allows to use a linguistically richer notion of context for the corpus-based creation of distributional approximations of language or semantics.

Early distributional models for English (Hindle, 1990; Grefenstette, 1992) used partial parsing – POS tagging and chunking, followed by a heuristic identification of grammatical relations from chunk order, while newer work for English (Lin, 1998a; Weeds, 2003) uses fast full parsing techniques.

For German, free word order and morphological variability create some difficulties (at least over English) which one should keep in mind and which to some extent make partial parsing techniques (as well as fast full parsing) more complicated than in English.

5.3.1 Part-of-speech tagging

Part-of-speech (POS) tagging consists in assigning parts of speech to words. After the first corpora with hand-disambiguated part-of-speech information became available, results using statistical part-of-speech tagging showed that it is feasible to treat such tagging as an autonomous task, with the benefit that statistics would take care of unknown words and the sequence model that underlies it would be less sensitive to complex syntactic constructions than a full parser would be.

Part-of-speech taggers usually combine information sources that allow the possible (or likely) tags tor the words in a text together with a model that allows to disambiguate the sequence; probabilistic models include Hidden Markov models, as in TnT (Brants, 2000), but also models that infer the tag distribution given the context using a decision tree (Schmid, 1995).

State-of-the art approaches using discriminative sequence tagging techniques reach accuracies of 97-98% for English and German (Shen *et al.*, 2007; Giesbrecht, 2008), which is sufficient for most purposes. Areas of concern for part-of-speech tagging are performance on out-of-domain data, but also types of errors that strongly impede the performance of subsequent modules relying on the POS tags.[12]

In German, sequence learning approaches yield systematic errors where locality assumptions do not hold, for example in difficult cases of relative/demonstrative pronouns, or in cases where a verb in clause-final position may either be finite or infinite:

(5.7) a. *Heute wollen* die Kinder baden gehen/VVINF.
today want the children swimming go[infinite].
Today, the children want to go swimming.

b. *Peter glaubt, dass* die Kinder baden gehen/VVFIN.
Peter believes that the children swimming go[finite].
Peter believes that the children will go swimming.

In many cases, it is possible to apply rule-based postcorrection to the output of a part-of-speech tagger using larger structural units – Müller and Ule (2002), for example, overcome some inaccuracies by a combined approach of topological field chunking and tag correction. Nevertheless, one should bear in mind that tagging errors do occur and that any approach that assumes POS tags (e.g., chunking or dependency parsing) may suffer from degraded performance as a result of error percolation.

5.3.2 Shallow and Incremental Parsing Approaches

A next step would be the chunking of noun phrases and verb groups; chunking is highly attractive for many applications since it is more robust than parsing, especially when only limited training data is available. In English, noun chunks (i.e., noun phrases without postmodifiers) are nonrecursive and can be found using simple sequence labeling techniques. In German, noun phrase chunks are not necessarily non-recursive, as adjective phrases can themselves embed argument noun phrases (and noun phrases generally can embed adjective phrases). This can create recursive structure within a chunk:

[12] Giesbrecht points out that part-of-speech taggers with excellent performance on news paper text from the same corpus they were trained on universally showed a severe performance degradation to about 90% accuracy when they were used to tag out-of-domain text from the World Wide Web.

(5.8) [$_{NP}$ Der [$_{NP}$ schönen, an [$_{NP}$ sonnigen Orten] gereiften Weinen]
[$_{NP}$ The [$_{NP}$ fine, at [$_{NP}$ sunny locations] matured wines]
niemals abgeneigte König] verbrachte den Großteil des Tages
never adverse king] spent the greater part the:GEN day:GEN
betrunken.
drunk.
The king, who was never adverse to fine wines matured at sunny locations, spent the greater part of the day in a drunken state.

Interaction between the boundaries of noun phrases and the argument structure of the verb are more frequent in German as well:

(5.9) a. ...dass [der König Johann] schläft.
...that [the King John] sleeps.
...that [King John] sleeps.
b. ...dass [der König] [Johann] schlägt.
...that [the King] [John] beats.
...that [the King] beats [John].

Additionally, the scope of conjunctions is often at least locally ambiguous, which makes clause (and field) chunking error-prone for sentences with clause coordination.

In applications where CPU time consumption is an important factor, methods that use deterministic local decisions, including partial parsing approaches, can be highly attractive since keeping track of ambiguities in the approaches based on beam search and/or dynamic programming is inherently more expensive, and usually one or two orders of magnitude slower.

For English, Curran and Moens' emulation of the SEXTANT partial parser (Curran and Moens, 2002b; Curran, 2004) – consisting of an off-the-shelf chunker and several postprocessing steps, has been shown to come near fast full parsers while being significantly faster and thus offers a compelling solution when enough text material is available. Due to German's less-configurational word order properties, such simple methods are unsuited for grammatical relation extraction. If similarly simple approaches were available for German with a fast implementation and acceptable quality, they would offer a compelling solution for the collection of grammatical relations (or other syntax-based information that can be used for distributional similarity measures) from large text corpora.

One of the few approaches to machine-learning based NP chunking of German is the CHUNKIE chunker by Skut and Brants (1998), which encodes local features of the syntax tree in tags, overcoming the difficulty of premodification by allowing more complexity in the tags as compared to commonly used schemes (cf. Ramshaw and Marcus 1995) which only encode chunk boundaries and are unable to cope with additional hierarchical structure. Other approaches, such as the one of Schmid and Schulte im Walde (2000), use a PCFG grammar for deriving chunks, which

does away with the promise of higher efficiency.[13]

The majority of the approaches to NP chunking instead use cascades of finite-state rules to annotate structure: both Müller and Ule (2002) and Kermes and Evert (2002) use several layers of rules to account for NP embedding into adjectival phrases. By inferring clause boundaries and looking at possible case assignments for noun chunks, Müller (2004a) is able to achieve a good approximation to predicate-argument structure in most cases.

Trushkina (2004) uses a rule-based approach that builds up phrase and clause chunks in combination with a morphological disambiguation based on a PCFG parser and subsequently uses disambiguation rules to assign grammatical functions. Due to the use of a PCFG parser for morphological disambiguation, this approach is quite slow, despite the rule-based part being executed on a fast (although proprietary) parsing engine.

All of these solutions are much more complex than partial parsers for English: Curran's emulation of the SEXTANT partial parser (Curran and Moens, 2002b) consists of an off-the-shelf chunker and several postprocessing steps, which are both rather simple to replace and/or optimize. Müller's approach of combining clause and noun phrase chunking with heuristics for argument assignment might offer both good performance and excellent speed if it were implemented in a more tightly-knit fashion without multiple steps of XML serialization and deserialization. However, it has to be said that the approach is far more complex than the simple pattern-matching techniques used by Curran and Moens.

Chunk parsing is not the only realization of the idea of deterministic parsing. By using a machine-learning classifier to predict the actions of a shift-reduce parser, it is possible to perform full parsing (in the sense that a complete parse is output) in a deterministic fashion, as proposed by Nivre (2003) for dependency parsing. While Nivre's original implementation does not excel in speed, due to the use of a memory-based-learning classifier,[14] an implementation by Kilian Foth which uses a simple statistical backoff model is part of the CDG distribution and reaches adequate accuracy when trained on a large amount of annotated data. The shift-reduce parser of Attardi and Ciaramita (2007), based on Maximum Entropy classification, is also claimed to perform at 200 sentences/second.[15]

5.3.3 Parsing by Global Optimization

In contrast to incremental approaches, in which mistakes cannot be discovered and later undone, approaches that are based on global optimization aim at finding a

[13] On the other hand, it should be noted that holistic approaches, while not being faster, can be, and sometimes are, more accurate: For English, Charniak's (2000) PCFG-based parser produces higher-quality chunks than the best available statistical chunkers, cf. Hollingshead *et al.* (2005).

[14] Memory-based learning and SVM classifiers exhibit a classification time that grows with training set size.

[15] `http://sites.google.com/site/desrparser/`

parse (according to a highly underspecified grammar) that maximizes a score that reflects the acceptability and/or probability of the parse.

As a result of combinatory effects, the number of parses for a given sentence shows explosive (i.e., exponential) growth with sentence length, which means that it is usually not feasible or desirable to score all sentences.

Therefore, global optimization in parsing uses either dynamic programming techniques (based on a scoring function that can be factored into parts, or equivalently, only uses *local* constraints) or uses approximate search – by assuming that dynamic programming conditions hold at least approximately (*beam search*) and/or by using heuristic optimization techniques; In contrast to greedy structure building, approximate search and heuristic optimization are, in principle, able to recover from local errors when disambiguating information is available non-locally.

Weighted Constraint Dependency Grammar One approach that has been shown to be very successful for parsing German is the Weighted Constraint Dependency Grammar, which has been realized in a broad-coverage parser for German (Foth *et al.*, 2004; Foth and Menzel, 2006) and allows the use of a wide range of graded constraints to assign scores to possible dependency parses, including nonprojectivities. The parser of Foth et al. reaches an unlabeled dependency accuracy of 91.9% (with 90.4% labeled accuracy for dependencies) with automatically assigned part-of-speech tags, using a grammar with hand-crafted constraints that are meant to reflect general grammaticality constraints such as agreement checking or ordering and attachment preferences.

The problem of finding an optimal parse in WCDG parsing is solved using a general-purpose optimization algorithm (Taboo search, with parsing-specific optimizations; see Foth and Menzel, 2003). While it is possible to get a parse at any point in the search process, the optimization time necessary to achieve the best results is usually longer than in dynamic-programming based algorithms.

PCFG Parsing Parsing based on probabilistic context-free grammars (PCFGs) uses a generative probabilistic model (assigning probabilities to each CFG production), which results in strong locality properties that allow polynomial-time parsing using a dynamic programming model.

These approaches occupy a kind of middle ground since they exhibit a decidedly non-linear relation between sentence length and parsing time (most usually, cubic or between quadratic and cubic), but cannot handle non-projective dependencies (in equivalent terms - crossing branches) or non-local constraints. Since non-projective dependencies (mostly relative clauses and infinitive clauses, which can be postposed) do not have a detrimental influence on the rest of the parse, the assumption of projectivity is rarely problematic for parsing.

Formally, a *context-free grammar* consists of a set of nonterminals N, a set of terminals T, a set of productions P which have one element from N on the left

hand side and a sequence of elememtns from $N \times T$ in the right hand side, as in

$$S \rightarrow NP\ VP$$

and a start symbol S, and can be summarized as a tuple (N, T, P, S). A context-free grammar generates a set of strings of terminals (i.e., a subset of N^*) by applying the productions as rules of a context-free rewriting system which, starting from S, rewrites a string of symbols by replacing one element of N that occurs as the left-hand-side of a production by the right-hand side of that production, until the string of symbols only contains terminals. Conversely, the rule applications that lead to a string of terminals can be summarized in a tree.

Context-free grammars can be parsed in $O(n^3)$ time with a dynamic programming approach where the parser keeps track of the derivability of a symbol over a consecutive span of input terminals; a derivable symbol-span pair such as $NP_{[1,4]}$ is called an *item*.

Probabilistic context-free grammars (PCFGs) assign a probability distribution $P(RHS|LHS)$ over the rules for each left-hand-side symbol, yielding a generative model that assigns a probability to each tree; this joint (generative) probability distribution over tree structure and generated sentence also yields a conditional probability distribution over the parse trees for a particular sentence that can be used to select a preferred parse tree for that sentence, which context-free grammars and other non-graded generative models do not allow to do in a straightforward fashion. Because probabilities are defined locally over rule applications, dynamic programming approaches for context-free parsing can easily be extended to compute the probability for the best parse tree under a derivation item (*viterbi probability*) or the best-scoring parse tree itself.

The most obvious way to derive a PCFG grammar from a treebank is to read off rules directly from the treebank structures and use the counts to derive model probabilities. Such an unmodified treebank grammar usually achieves much larger coverage than a hand-crafted phrase structure grammar, but the parse quality is rather limited when compared to approaches that take into account more information than is directly encoded into treebank symbols such as lexicalized PCFG parses or approaches that derive PCFGs from a transformed treebank which contains more linguistic information in the node labels (see below).

For English, Charniak (1996) reports PARSEVAL results of 78% labeled F-measure for an unmodified treebank grammar, compared to about 86%-92% that can be reached using state-of-the-art approaches including treebank transformations, lexicalization, and more elaborate contexts (e.g., Klein and Manning, 2003a; Schmid, 2006 for treebank transformations, Collins, 1997; Charniak, 1997; Klein and Manning, 2003b for head-lexicalized PCFG parsing, or Charniak and Johnson, 2005; Huang, 2008 for more complicated approaches). For the German NeGra/Tiger treebanks, a plain treebank grammar reaches about 66%-70% compared to 76%-81% for more elaborate approaches such as the treebank transformation with smoothing of Dubey (2005), or the latent-variable approach of Petrov and

Klein (2008b) which was the best-performing entry at the shared task in the ACL 2008 Parsing German workshop (Kübler, 2008).

Head-Lexicalized PCFGs Head-lexicalized PCFGs augment a PCFG backbone by assigning a head constituent to each production, which allows the percolation of terminals as heads of constituents and leads to parse items such as $NP_{[i,j,\text{manager}]}$. A generative probabilistic model can then condition the probability distribution of rules, or that of generated lexical heads for the dependents (i.e., non-head constituents in a rule) on the lexical head of the head constituent of a rule. Because of sparse data problems, successful approaches for head-lexicalized PCFGs use sophisticated techniques for the smoothing of production probabilities and grammar rules which ultimately need relatively careful engineering both on the side of the probability model itself and on the side of the implementation.

Schulte im Walde (2000) presents a grammar for German verb-last clauses that uses linguistic features to derive the case of noun phrases from the morphology of the adjectives and determiners, starting with a lexicon generated with the help of DMOR (Schiller and Stöckert, 1995), a morphological analyser for German, and subcategorizing verbs for all frame types that they occur with. It has to be said that while the grammar was only trained in an unsupervised fashion, had limited coverage (about 92%) and only covered verb-last clauses, it is notable as an example of a carefully engineered PCFG grammar which makes good use of linguistic information.

Dubey and Keller (2003) implemented the head-lexicalized parsing models of Collins (1997) and Carroll and Rooth (1998) for German. Collins' model 1, on which Dubey and Keller base their experiment, uses a markovised phrase structure grammar, predicts arguments and adjuncts based on information from the head constituent and the distance to it, including bilexical statistics predicting headword-dependent relations. The model of Caroll and Rooth marks the heads of phrases, but, instead of using markovization, uses equivalences between the head-dependent relationships in different rules to prevent an explosion of the feature space.

Dubey and Keller compared an unlexicalized PCFG baseline model with a model based on Carroll and Rooth as well as their reimplementation of Collins' model 1, and they found that the unlexicalized model outperformed both Carroll and Rooth's model and their reimplementation of Collins' parser. This finding, lexicalization hurting performance, is at odds with corresponding findings for English (cf. Magerman 1995, Collins 1997), where lexicalized PCFG models outperform their non-lexicalized counterparts by as much as 10%, also noting that the effect does not seem to be due to sparse data since both the baseline and the C&R model achieve relatively good performance when trained on only 10% of the training data. They also observe that their reimplementation of Collins' model 1 is not at all affected by training set size, delivering similar performance at 10% and 100% of the training set size.

Dubey and Keller then modified their statistical model to include condition-

ing on the sister node, including dependencies on the sister category, head word, and head tag, essentially replacing the zeroth-order Markov assumption of Collins' model 1 with a first-order Markov assumption.[16] The resulting model resulted in an improvement over both the earlier model and the baseline.

Kübler *et al.* (2006) adapted the Stanford parser (Klein and Manning, 2003b), which already had a parsing model for German based on the NeGra corpus, to the TüBa-D/Z. The Stanford parser constructs parses using a factored model consisting of a PCFG-based model and a dependency-based model, where the PCFG model's items include annotated heads but lexicalization is realized in the dependency-based model. The factoring into PCFG-based and dependency-based models allows the use of exact search heuristics for both partial models, yielding a substantial improvement in speed while still guaranteeing exact inference. Kübler *et al.* compare the NeGra model of the Stanford parser in a variety of settings to the TüBa-D/Z model; besides the large difference in PARSEVAL bracketing score, they also find that using similar settings of second-order markovization and parent annotation, the TüBa-D/Z model produces more accurate results for subject and accusative/dative object NPs, which would be contained in equal number and therefore yield a more meaningful comparison than PARSEVAL figures.[17] The comprehensive evaluation of Kübler *et al.* (2008) suggests that the TüBa-D/Z model is slightly better at constructing the overall structure (slightly higher accuracy in the unlabeled dependency evaluation), whereas it performs worse than the Tiger-trained model in terms of dependency label accuracy.

Since German is a non-configurational language, grammatical relations are not evident from pure phrase structure parses, which means that grammatical function labels are needed in addition to phrase structure. The simplest and most common way of achieving this is by attaching grammatical functions to the node labels. However, this creates sparse data problems and suffers from the fact that the information that is available locally is not sufficient. The submission of Rafferty and Manning (2008) to the shared task ACL 2008 Parsing German Workshop (Kübler, 2008) demonstrates these findings: Including the edge labels results in a drop of constituent accuracy that is especially large on the Tiger treebank, where the absence of unary projection, and flat annotation scheme already result in sparse data. Furthermore, results for constituent evaluation including grammatical function labels were worse than those of other shared task participants.

[16] Note that Collins' model 2 includes a conditioning context of previously occurred arguments, which weakens the zeroth-order Markov assumption. Collins (1999) notes that the distance feature in his model 1 already eliminated some of the non-beneficial effects of the zeroth-order Markov assumption.

[17] A look at the exploration of parameter settings in Rafferty and Manning (2008) reveals that these settings - parent annotation and second-order markovization - are near-optimal for TüBa-D/Z whereas the parent annotation causes a sparse data problem in the Tiger annotation scheme when including grammatical function labels.

Annotated Treebank Grammars In work for English, Klein and Manning (2003a) were able to improve an unlexicalized PCFG model by using markovization and including additional linguistically motivated information in the labels. Without any modification in the treebank or in the model of unlexicalized PCFG parsing, Klein and Manning realized an improvement from 72.6% to 87.0%, a score that rivals that of earlier lexicalized models. Inspired by such results, Schiehlen (2004a) enriches trees from the NeGra treebank with additional linguistic information. Schiehlen's experiments with markovization indicated that while markovization can be used to improve the constituent-based score, it does not improve the dependency score. In the final version of annotated PCFG, Schiehlen uses deepening the flat structure of the NeGra treebank and adding case distinctions, subcategorizing dependent clauses into relative, complementizer-introduced subordinating and other phrases, and annotating the verb form, as well as named entity information (using an independent module that was based on gazetteers but also tuned to the training set). The resulting accuracy, measured in F-measure over (unlabeled) dependencies, improves from 78% for the baseline to 84% for the full model on the evaluation data set. Schiehlen also notes that imposing the POS tags of an independent part-of-speech tagger, as done by Dubey and Keller, works less well than using a morphological analyzer (Lezius *et al.*, 2000) to predict possible tags.

In a later paper, Dubey (2005) uses case marking and clause subcategorization in conjunction with smoothed second-order markovization. Dubey uses smoothed second-order markovization in conjunction with beam search, which yields a robust and very noticeable improvement (from 72.6% to 76.3% consituent F-measure in the case of his best-performing grammar). Dubey also notes that while smoothing is useful across all grammar variants, the different smoothing algorithms react differently to the various grammar transformations.

Versley and Rehbein (2009) combine an annotated grammar approach with lexicalization based on labeled dependencies in a feature-rich discriminative model including unsupervised PP attachment and selectional preferences. This approach is able to identify grammatical functions much more accurately than the Stanford parser while also offering slightly better results for the identification of constituents (i.e., ignoring grammatical functions).

Dynamic Programming Approaches to Dependency Parsing Several approaches exist for unlabeled dependency parsing, which allow for faster lexicalized parsing than head-lexicalized PCFG approaches due to stringer independence assumptions. The minimum spanning tree-based dependency parsing approach of McDonald *et al.* (2005) allows directly parsing non-projective structures with global optimization of a model based on scores for single edges in $O(n^2)$ time. The subsequent model of McDonald (2006), which uses an $O(n^3)$ time complexity algorithm similar to the one proposed by Eisner and Satta (1999), allows scores for pairs of same-parent edges, but disallows projective structures. Non-projective

parsing can be achieved in the second model by getting the optimal projective structure and then doing a hill-climbing search reattaching edges.

State-Split Grammars The locality assumption that is inherent to PCFG grammars derived from treebanks has been shown to be both too loose (in the sense that treebank productions must be split into further parts using markovization to ensure that unseen productions can be handled more gracefully) and too strong (in the sense that a pure treebank grammar does not contain enough information to ensure accurate parses). While annotated treebank grammars solve this problem – important linguistic information is added to the trees and can be fruitfully used in parsing, improving the accuracy and resulting in more plausible parses – it would be desirable to have an approach that automatically learns such transformations from the training data.

Such state-split approaches, which split the original labels from the treebank into multiple distinct values use the data in the original treebank to evaluate the usefulness of splits using goodness-of-fit statistics instead of relying on linguistic intuition. (Such an approach may yield symbols such as *NP-23* and *NP-42* instead of simply *NP*, with the potential benefit that important differences such as the one between relative noun phrases and ordinary noun phrases can be exploited by the parser despite both kinds receiving the same node label *NP* in the original treebank),

The Berkeley parser (Petrov *et al.*, 2006; Petrov and Klein, 2008a), automatically acquires subcategorization by splitting states and using an EM-based algorithm to refine rules; splits are reversed again when the performance on a separate development set does not increase. This is similar to work by Ule (2003) in that states are split, the grammar optimized using the EM algorithm and splits are possibly undone when they turn out to be undesirable. While Ule uses distributional considerations to motivate splits and propose merging of similarly-behaving subsymbols, Petrov *et al.* use a simpler heuristic of proposing a binary split for every symbol, and undoing a split whenever it does not result in a large enough decrease of the likelihood of the held-out data.

In contrast to Ule, Petrov *et al.* follow Matsuzaki *et al.* (2005) in approximating the probability sum over different subsymbol assignments for one coarse-grammar tree rather than taking the most-probable fine parse. Matsuzaki *et al.*'s findings, as well as those of Petrov *et al.* suggest that this so-called *max-rule* strategy yields an improvement of as much as 2% in terms of Parseval F-measure.

As Ule only evaluates his approach on the task of topological field chunking for German, where published research using similar methods and evaluation is scarce, it is not immediately clear if his particular choice of training and inference procedures would perform better than Petrov's or the ones by Matsuzaki *et al.* (2005) or Prescher (2005), who uses Viterbi decoding and a different treebank transformation in his latent-variables approach and achieves a smaller improvement than Matsuzaki *et al.*

The results of experiments by Cheung and Penn (2009) suggest that the Berkeley parser's performance in indentifying topological fields is superior to Ule's results: They find an F-measure of 93.35% for the Berkeley parser, against 89.90% for Ule's approach and 85% for a baseline PCFG when using automatically identified part-of-speech tags.[18]

An advantage to the implementation of Petrov *et al.* (2006) over earlier work is also the use of coarse-to-fine parsing, which prevents the parsing time from growing too much with grammar size.

Reranking approaches A different method to overcome the locality assumptions of a treebank PCFG (or indeed any PCFG) is the approach of reranking (Collins, 2000), in which the highest-scoring parses from a PCFG parser are taken and a more elaborate model, which does not need to abide the locality assumptions of a PCFG, is applied to obtain the new score.

Zinsmeister (2008) uses subtree boosting (Kudo and Matsumoto, 2004), a machine learning technique that assigns weights to tree fragments that can encompass multiple layers in the tree. Zinsmeister interprets the ranker output as a binary classification, and as a result her reranking procedure only makes a difference for very few sentences, where, however, the reranker's choice is a vast improvement in parse quality. Seen globally, the improvement is very limited as the vast majority of sentences are not changed at all from the 1-best hypothesis. (Zinsmeister evaluates labeled bracketing measures on an enriched version of treebank trees, which means that her results cannot be compared to any other approach).

Kübler *et al.* (2009) use the reranking mechanism of Collins and Koo (2005) on top of an unmodified unlexicalized treebank PCFG and show that reranking gives significant improvements above their baseline. Additionally, they show that using linguistically sensible heuristics to increase the diversity in the list that is given to the reranker by enforcing different possible coordination scopes, it is possible to reach further improvements. Although the improvements reported by Kübler *et al.* are not incremental (their baseline is much below current state-of-the-art results such as those reported in Kübler, 2008, to the point where the substantially improved results with the more advanced techniques are still below those of more standard head-lexicalized PCFGs), they show that reranking approaches are useful even in such a setting, and can be improved further by avoiding search errors through better-informed sampling strategies.

[18] Ule uses TnT to assign part-of-speech tags, whereas Cheung and Penn use the Berkeley parser's unknown word model. Results using gold part-of-speech tags – a constituent F-measure of 95.15% for Cheung and Penn's results using the Berkeley parser and an F-measure of 91.98% for Ule's approach – show a similar tendency.

5.4 Summary

Coreference resolution, especially if aiming at more ambitious targets such as coreferent bridging, depends on a wealth of information sources to identify necessary features both during pre-processing of text and during resolution. As a result, the difference in available resources between English and German – some resources that have been freely available for English are either not or only recently available to the research community at large, and are, as a result of more limited choice, more difficult to integrate.

In this chapter, I have reviewed the available resources for German that can be used to implement semantic information sources such as those used by the approaches presented in section 4.3. It is against this backdrop that I will present the information sources and the distributional similarity models in the next chapter, pointing out necessary differences to such an approach for English. German's flexible word order makes straightforward solutions, including partial parsing, much less attractive. The application of unsupervised learning techniques for German also requires solutions for consistent morphological analysis, lemmatization, and compound splitting.

Chapter 6

Semantic Compatibility for Coreference

This chapter will focus more closely on various techniques to resolve coreferent bridging with good recall, and present an implementation of both existing techniques and improvements over them in the context of German newspaper text.

In the review of linguistic approaches in section 2.2, I have outlined the importance of *logical relations* such as *subsumption* and *compatibility* as the logical relations that constrain the antecedent choice for an anaphoric definite description in cases where lexical identity or synonymy cannot be used to identify an antecedent.

In contrast to such logical relations, *association* and *similarity measures* are most commonly symmetric and are continuous-valued. As discussed in section 4.3, it is useful to distinguish among two kinds of these measures: One is *semantic similarity* (which may intuitively correspond to neighbourhood in a taxonomy, or the proportion of shared distinctive features), and the other is thematic relatedness (which may link concepts from different ontological kinds, for example *bus* and *driver*).

We can motivate an intuition that two mentions may be coreferent based on one of the following relations which may hold between their heads:

- *near-synonymy* in the case where both terms can be used interchangeably in most contexts
- *hyperonymy* for a subsumption relation between two concept terms
- *instance* for a link between a name (i.e., a single instance) and a concept describing it
- *compatibility* for two terms that are incomparable as to their logical content, but not contradictory

Similarity measures (either taxonomic similarity measures derived from a wordnet, or distributional similarity measures induced from a corpus) are undi-

rected. In the categorization above, they could be seen either as targeting near-synonymy relations (for terms that are often, but not always, used synonymously) or as targeting compatibility (for terms that can conceivably apply to the same thing but mean different things, such as *book* and *birthday present* or *chairman* and *drinker*).

A successful combination approach would combine indicators for the different types, to provide high-precision coverage for those cases where stronger assumptions can be made, and also to help cover those cases where the stronger assumptions do not hold. To provide an example using the hand-crafted resources presented in the last chapter, synonymy and hyperonymy in GermaNet can be checked directly, whereas the existence of a short path in GermaNet can be indicative of either near-synonymy or at least indicate the existence of a compatible supertype.

Section 6.1 reviews techniques for implementing distributional similarity and association measures and motivates a selection of three different measures that can be implemented in German.

Section 6.2 addresses the more immediate question of implementing distributional similarity measures and the other approaches mentioned earlier. Based on the overview on the available parsing techniques that could be applied to German (section 5.3.3), I present the parser that has been used to provide analyses for the taz corpus of newspaper text (section 6.2.1), and discuss other components such as the extraction of grammatical relations (6.2.2) and the treatment of compounds (6.2.3), which is necessary to achieve good results in German.

Section 6.3.1 will introduce an approach using pattern search on the World Wide Web to acquire instance relations (such as *Bremen – city*), harnessing the advantage in size of the Web indexed by search engines compared to the relatively smaller corpora that researchers commonly have access to.

Finally, in section 6.4, I will present a system that allows the resolution of anaphoric definite noun phrases to a suitable antecedent, exploiting all the knowledge sources discussed before. This system makes it possible to compare the advantages and drawbacks to the approaches discussed in the earlier section, and will be instrumental for creating a highly effective combination-based approach to the resolution of definite NP anaphora.

6.1 Corpus-based Similarity and Association Measures

As mentioned in the introduction to this chapter, it is useful to keep in mind several distinctions that have an influence on the range of potential applications for a statistical measure for word pairs:

The first is whether such a measure is a (symmetric) measure of (topical or semantic) similarity, or whether it serves to detect a relation (such as hyperonymy) between two nouns. While instance relations (hyperonymy in GermaNet) between names and common nouns (*Corsica–island*) are very useful, similarity relations

between names (such as *Corsica–Elba*, which are both mediterranean islands) are not indicative of coreference.

The second distinction is between semantic similarity on one hand and topic similarity (general association, or relatedness) on the other hand – in general, relatedness measures also capture associated pairs that are not similar semantically such as *door* and *house* in the following example (5.2c):

(6.1) John walked towards [$_1$ the house].

(6.2) a. [$_1$ The building] was illuminated.
b. [$_1$ The manor] was guarded by dogs.
c. [$_2$ The door] was open.

Typical cases of coreference include cases like 1,2a (*the house–the building*: hyperonym) or 1,2b (*the house–the manor*: non-synonymous term which is not subsumed by the antecedent, in this case, a hyponym). The discourse in 1,2c is an example of associative bridging between the NP *"the door"* and its antecedent to *"the house"*; it is inferred that the door must be part of the house mentioned earlier (since doors are typically part of a house), which is *not* compatible with coreferent bridging, but is also ranked highly by association measures.

6.1.1 Distributional Similarity Measures

It is well-established that human learning of lexical items beyond a certain point is driven by considering the *contexts* in which a word occurs, and it has been confirmed by McDonald and Ramscar (2001) that few occurences of a word in informative contexts suffice to influence similarity judgements for marginally known words.

The basic idea of distributional similarity measures is to provide a statistical approximation to lexical semantics by recording the contexts in which a word occurs (usually representing them by the lexical items that occur in the syntagmatic neighbourhood of that word), and then using the distributions of co-occurring items, either looking at specific items or using similarity measures on these distributions, as a proxy for word meaning, or at least for word usage in a specific text, as in Philips (1985).

For *collocates*, words co-occuring significantly more often with a given word than it would be expected for random text, Church and Hanks (1990) demonstrate a correlation with psychological findings about priming and/or free association, and point out that collocates sometimes, but not always, can be explained by some lexical or syntactic relationship between the words.

One can roughly classify approaches to use collocates into two categories: the approaches of Philips (1985) and Church and Hanks (1990) find words that co-occur in the *same* contexts (**first-order collocates**) which means that they are highly associated, but the relation between them will rarely be one of identity and/or synonymy since synonyms co-occur only very rarely, which means that collocates will be dominated by other lexical relations.

Using collocates as *features* in a similarity measure that reports words as similar if they frequently have the same collocates will yield a significant portion of synonyms and near-synonyms among the terms with similar collocate distribution (which are sometimes called **second-order collocates**): Consider *cosmonaut* and *astronaut*, two words which both denote crew members in space travel. Even though they might be rather unlikely to co-occur, they will be used in very similar contexts, yielding a high similarity value between both words.

It is however important to note that this similarity relation is generally *not* an iteration of collocations in the most apparent sense: If we have a matrix A that maps words to its collocates, a naïve notion of second-order collocations would suggest $AA = A^2$ as the collocation measure, whereas our similarity measure is based on the symmetric definite matrix $A^T A$. The two matrices are of course equal if the collocation relation is symmetric, i.e., $A = A^T$ (e.g. in the case of a symmetric window and target and feature words being the same).

Let us use an example to illustrate the general procedure of getting from collocate words to a distributional similarity measure. In the subsequent paragraphs, each part of the process will then be discussed in more detail.

For the example, we take some words *eagle,penguin,kettle* with co-occurrence features *feather,fly,tea,black*. Our (yet-to-be-specified) association measure might yield the following weights:

	feather	fly	tea	black
eagle	2	3	0	1
penguin	2	0	0	1
kettle	0	0	2	1

We can then use the weight vectors to represent the words and use a similarity measure over vectors to calculate the similarity between two words.

A fairly standard one is the cosine similarity measure, which is based on the fact that the dot-product of two vectors is the product of their length times the cosine of the angle between them. Cosine similarity and other similarity measures will be discussed in greater detail in pages 157 and following; For vectors pointing in the same direction, the cosine is 1, for vectors pointing in orthogonal directions, the cosine is 0 and for vectors pointing in opposite directions, it is -1.

Using the cosine similarity measure $\mathrm{sim}_{\cos}(w_1, w_2) = \left\langle \frac{w_1}{\|w_1\|}, \frac{w_2}{\|w_2\|} \right\rangle$ on the rows of the above matrix, we could then calculate the similarity of *eagle* and *penguin* as $\frac{2 \cdot 2 + 1 \cdot 1}{\sqrt{14} \cdot \sqrt{5}} \approx 0.60$, and the similarity of *eagle* and *kettle* as $\frac{1 \cdot 1}{\sqrt{14} \cdot \sqrt{5}} \approx 0.12$.

The main differences between approaches to distributional similarity measures can be roughly broken down along the axes sketched in this example: first, how contexts of a word are described, usually in terms of collocates (*feature extraction*); second, how the features characterizing the features are weighted to yield a numerical vector that can be used as a proxy for word meaning (*feature weighting*), and third, how the similarity between the numerical vectors is measured (*vector similarity*). Finally, some approaches use the vectors obtained through such means and

apply a transformation such as singular value decomposition to improve the quality of the obtained similarity measure (*postprocessing and dimensionality reduction*).

Feature Extraction It is possible to group the approaches to use collocate features into two main areas:

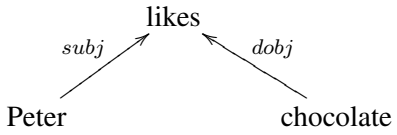
- relation-free methods aim to directly use vectors of co-occurring words as a representation without distinguishing the relation between the target word and its collocates. Thus, related terms such as *doctor*, *hospital* and *treatment*, which share many collocates, would be assigned a high similarity value.
- relation-based methods use collocate words together with grammatical relations, so that one noun being a frequent subject and another being a frequent object of a given word would not increase their similarity score – as an example, a context like *the doctor treats the patient* would not contribute to the estimated similarity between *doctor* and *patient*.

Relation-based methods would be expected to yield more informative weightings than relation-free methods (since *direct object of eat* is more specific than simply *co-occurs with eat*, cf. figure 6.1), but they could also be seen to make less effective use of the available data since relation-based approaches only consider a small set of interesting relations whereas relation-based methods could more easily capture more indirect associations (as an example, consider the relationship between *goat* and *grass*, which is not expressed by a single grammatical relation, yet would be seen as an informative collocation).

In a similar vein to the distinction made between relation-based and relation-free approaches, Evert (2005) uses a distinction of *positional cooccurrences* versus *relational cooccurrences*, where positional cooccurrences are simply based on surface distance whereas relational cooccurrences that are based on a particular grammatical relation. Evert's distinction is however not applicable to the relation-free approach of Padó and Lapata (2003) since their approach uses explicit grammatical relations to derive a neighbourhood relationship. The fact that Padó and Lapata's approach uses grammatical relations would be a characteristic feature for Evert's relational cooccurrences, but their approach uses only the cooccuring words without distinguishing between different relations, and gives results that are closer to the thematic similarity of positional cooccurrences.

Among relation-free methods, **window-based** feature extraction is the most popular alternative: given a span of text, simply extract all words that are within a certain distance to the target word; this can range from one content word left or right of the target word, to multi-word windows, to using the whole paragraph or the whole text as neighbourhood.

Because window-based feature extraction is computationally cheap and does not require preprocessing of any kind, it was used not only by early approaches to extracting collocates such as those by Philips (1985) or Church and Hanks



Lin98: (person) ⇒ ↑SUBJ : like
chocolate ⇒ ↑DOBJ : like

PL03: (person) ⇒ like(1.0), chocolate(0.5)
chocolate ⇒ like(1.0), (person) (0.5)

Figure 6.1: Syntax-based collocate features: relation-based and relation-free

(1990). It also forms the basis of later approaches for discovering second-order co-occurrences (i.e., thematic or semantic similarity) such as Latent Semantic Analysis (Dumais *et al.*, 1988; Deerwester *et al.*, 1990), the distributional model of Schütze (1992), or the Hyperspace Analogue to Language model of Lund *et al.* (1995). Within the window-based approach, variation exists regarding the exact way of measuring the window – Latent Semantic Analysis proposes the use of paragraph or document windows, Schütze uses a 1000-character window around the target, whereas Lund *et al.* used a window of ten content words. It is common to limit the words used as features to a list of most-frequent words to avoid data sparseness; Lund *et al.* use the 200 words which showed the greatest variance in co-occurrence counts across different targets among the 70 000 most frequent words in the corpus.

Lund *et al.* found that the vector-space model resulting from their approach was able to reproduce coarse semantic distinctions such as between animals, countries, and body parts. In further results reported in the same article, they showed that the model was a good predictor for semantic priming effects when the items were semantically similar. In the light of previous experiments on semantic priming, which suggested that both semantic similarity and association were necessary to get priming-related effects, their model can be thought of as relatively accurately modeling the association part.

Schütze's characterization of his model also emphasizes the presence of topically, rather than semantically similar words, saying that neighbour sets were more clearly delineated for words with a clear topic connection such as *Makeup*, *mediating* and *litographs* than for words like *keeping* and *absolutely*.

An obvious alternative to window-based approaches are those that target a set of specific **grammatical relations** such as subject-verb and object-verb relations. In these approaches, the relation between target word and collocate is made explicit by using tuple of the relation and the co-occurring word as feature, which yields features that are more closely related to actual properties (such as: "can be eaten" or "has wings") than in relation-free methods.

To create these grammatical relations, of course, further processing is necessary: Hindle (1990) used a deterministic partial parser (Fidditch, cf. Hindle, 1983) to extract subject-object and subject-verb pairs, whereas Grefenstette (1992) used a larger set of relations (premodifying adjectives and nouns, PP postmodification, as well as subjects and direct and indirect objects) using a chunker and a pattern-based recognition of grammatical relations, whereas later work such as that of Lin (1998a) or Weeds (2003) use generic full parsers – Lin's own Minipar in his case (Lin, 1998b), and the RASP parser (Briscoe and Caroll, 2002) in the case of Weeds.

Quite obviously, there is a tradeoff between the precision of full parsers and the greater speed of window-based approaches; Curran and Moens (2002b), and later Curran (2004) note that due to the limited availability of suitable text, fast full parsers such as RASP or MiniPar would still be preferable when the computing time is not strongly limited (Curran notes that Minipar takes four days for processing 150 million words, which would be an acceptable time in any case).

Padó and Lapata (2003; 2007) define a word-based context model using **syntactic neighbourhood** instead of surface neighbourhood, defined over paths of dependency relations. They report that their model works significantly better than a simpler window-based approach on a set of tasks, and even slightly outperforms a model based on grammatical relations on the task of determining predominant word senses.

It is also possible to use linguistic structure in other ways than Padó and Lapata's approach of using grammatical relations in relation-free approaches: In particular, it is also possible to use a more greater variety of relations that is not limited to grammatical ones: simple **semantic patterns** that indicate parthood, hyperonymy, or other relations can also be used as features, provided that the text source is large enough to allow the retrieval of such patterns with sufficient recall.

For example, Almuhareb and Poesio (2005a) target attributes and attribute values of nouns (for example, *dogs* have a *color*, which is *brown*) using adjective-noun and part-of patterns. Katrenko and Adriaans (2008) use patterns targeted at the acquisition of qualia structures using search on the World Wide Web, and Versley (2008a) uses instances of Hearst's hyperonymy patterns on Google's data set of 5-gram counts from a large sample of the World Wide Web[1] to complement higher-recall syntactic patterns used on the UK-WaC corpus.

Feature Weighting Constructing vectors from the raw frequency counts of co-occurring elements – be it from a window-based approach or from something more elaborate – is not an ideal method since these frequencies are essentially incommensurable: frequent (but not necessarily informative) features will swamp the infrequent ones, and target words with very different frequencies may have vectors that are not at all close or similar to each other.

[1] Thorsten Brants, Alex Franz (2006): Web 1T 5-gram Version 1; LDC2006T13

In the context of information retrieval using Latent Semantic Indexing (LSI), Dumais (1991) discusses methods to determine the vectors for target words, which she claims improve retrieval performance with LSI by as much as 40%; further work in distributional similarity measures, such as Grefenstette (1992) or Rapp (2003) have used some of the measures proposed.

Dumais separates the weighting strategy into two parts: a *local weighting* part that reflects the association of features with the document, and a *global weighting* part that emphasizes or deemphasizes features (with the goal of putting an emphasis on more informative features).

In her study, Dumais uses the following local weightings:

- Term frequency (n_{ij})
- Binary (1 if word i appears in the document, 0 otherwise)
- Logarithmic ($\log(n_{ij} + 1)$)

together with the following global weightings:

- Normalize: $\frac{1}{\sqrt{\sum_j n_{ij}^2}}$
  (i.e., dividing by the L_2-norm of the weights for one feature)
- Idf: $\log_2\left(\frac{ndocs}{\mathrm{df}_i}\right) + 1$
  (inverse document frequency, log-transformed)
- GfIdf: $\frac{n_i}{\mathrm{df}_i}$
  (Global Frequncy times Inverse Document Frequency – the average number of occurrences in those documents where it occurs)
- Entropy: $1 - \frac{-\sum_j p_{ij} \log(p_{ij})}{\log ndocs}$
  where the entropy for a term is divided by the maximal value (which would be $\log ndocs$ for a term that occurs equally frequent in each document) and then subtracted from 1 to yield a value that is 0 for terms that occur equally in every document and 1 for terms that only occur in one document.

where n_{ij} is the frequency of term i in document j, n_i is the total frequency of term i in the document collection, and df_i is the number of documents a term occurs in, whereas $p_{ij} = \frac{n_{ij}}{n_i}$ is the proportion of total occurrences that occur in one document, and $ndocs$ is the total number of documents.

The global weighting schemes are chosen as to lower the importance of terms that are either very frequent, or that predictably occur in many of the documents (whereas interesting terms tend to either be rare altogether or occur in a bursty fashion, with documents typically containing several occurrences).

In Dumais' study, the entropy-based measures (both logarithmic and normal counts) worked best, followed by tf-idf (i.e., term frequency with global weighting by inverse document frequency), whereas dividing feature weights by the L_2 norm for a feature performed worst of all. In parallel with this, Grefenstette (1992) reports that of the term weightings tried in his experiments, the log-entropy weighting also worked best.

In contrast to these more ad-hoc weighting techniques, other measures that have been proposed are founded on the intuition that the weight should either reflect the ratio of actual to expected counts (according to some background distribution), or based on statistical significance (which gives an estimate for how unlikely it is that the background distribution would have generated these counts).

The **pointwise mutual information** measure uses the ratio between expected and actual occurrences of a feature to get weights from these counts that determine the (estimated) degree of association between subject and verb:

$$w_{\text{mi+}}(x, y) = \begin{cases} 0 & \text{if } p(x,y) < p(x)p(y) \\ \log \frac{p(x,y)}{p(x)p(y)} & \text{otherwise} \end{cases}$$

(where x and y again designate target word and feature, as do i and j in the above examples).

Because the pointwise mutual information measure gives large negative numbers when $p(x, y)$ is small in relation to $p(x)p(y)$, or becomes infinite in the case where no pair was observed (i.e., n_{xy} is zero), it is common practice to cut off negative values (Lin, 1998a).

A third group of measures is related to **statistical tests** that compares the relation between a null hypothesis (that a feature y occurs as many times as it would be if it were distributed randomly) and the empirical distribution of the features.

The basis for these statistical tests is the number of occurrences and non-occurrences of x (one specific target word) and y (one feature) in the target word – feature pairs that have been extracted. Thus, for a given target word / feature pair, one can get the following counts:

n_{xy}	$n_{\bar{x}y}$	n_y
$n_{x\bar{y}}$	$n_{\bar{x}\bar{y}}$	$n_{\bar{y}}$
n_x	$n_{\bar{x}}$	

Looking just at the *marginal* distribution, we can compute *expected counts* $\hat{n}_{xy} = n \cdot p(x) \cdot p(y) = n_x \cdot p(y) = n_y \cdot p(x)$ for $xy, x\bar{y}, \bar{x}y, \bar{x}\bar{y}$.

Two standard empirical tests, the *t-test* and the *Pearson's χ^2 test*, are popular means to assess whether the difference between the expected and observed counts is likely to occur by chance.

The t-test divides the actual difference between expected and observed counts by an estimate for the standard deviation (i.e., shifting and scaling the differences

so that they correspond to a distribution with a mean of 0 and a variance of 1):

$$w_t(x,y) = \frac{n_{xy} - \hat{n}_{xy}}{\sqrt{n_{xy}(1 - p_{xy})}}$$

Pearson's χ^2-test has been used as a collocation statistic by Church and Gale (1991) and others, and is an estimate for the Euclidean distance of the counts (seen in 2-dimensional space) to the expected counts; the intuition behind this is that in the case where x and y are actually correlated, the observed counts would deviate further from the expected counts than would be expected under the null hypothesis (i.e., independence):

$$w_{\text{Pearson}} = \sum_{\substack{i\in\{x,\bar{x}\}\\ j\in\{y,\bar{y}\}}} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Both the t-test and Pearson's χ^2 test are based on the assumption that the counts are distributed according to a normal distribution, which does not hold in the case that is most common for distributional similarity due to the fact that the actual counts (especially of target-feature collocations for non-frequent words) are too small.

Dunning (1993) contends that, for low-frequency words, the normality assumption does not hold even approximately. He proposes the use of a likelihood ratio test where the likelihood of the data under (i) the assumption of independent sampling and (ii) a joint distribution for the x and y outcomes is assumed, under the general assumption that all counts are binomially distributed. The resulting statistic is again χ^2-distributed:

$$w_{LL} = \log \frac{(p_x p_{\bar{x}} p_y p_{\bar{y}})^{n_y}}{(p_{xy} p_{\bar{x}\bar{y}})^{n_{xy}} (p_{x\bar{y}} p_{\bar{x}y})^{n_{\bar{x}y}}}$$

Even though it is more reliable on rare items, Dunning's likelihood ratio statistic shares two weaknesses with Pearson's chi-square test that lessen its atractiveness as a weighting function for distributional similarity measures:

- It is a two-sided test, which means that dissociated terms (i.e., those that occur together significantly less than expected) will also get a positive score. This can be remedied by using 0 for those items where the actual count is less than the expected count.

- Like the t-test statistic, they are estimates of *statistical* significance rather than of the actual strength. This means that even among features that are sufficiently frequent, features that show a similarly strong association but are more frequent will be preferred. In medical studies, this difference is described as the one between statistical significance (which grows as the sample size grows) and clinical significance (the actual improvement that comes from the treatment, which does not grow with sample size).

> While factoring in frequency is a desirable property for the extraction of multiword units (i.e., simple collocations), it may be a hindrance to (some or all) vector similarity methods when extremely frequent features drown everything else or when the vectors for rare and frequent words have different scales.

Johnson (2001) proposes a statistic that combines a more cautious approach to lower-frequency associations with a behavior closer to the mutual information statistic for higher-frequency associations, namely calculating a confidence interval for the odds ratio

$$\frac{p_{xy}}{p_{\bar{x}y}} : \frac{p_{x\bar{y}}}{p_{\bar{x}\bar{y}}} = \frac{p_{xy}p_{\bar{x}\bar{y}}}{p_{x\bar{y}}p_{\bar{x}y}}$$

(according to an error bound that is determined by the user) and using that error interval to choose a more conservative (i.e., lower) estimate of that statistic.

As a result, the measure prefers strong collocations that are non-rare (e.g., "Los Angeles" in the case of bigrams) over weaker collocations that are very frequent ("of the" or "in the"). However, exact determination of this interval (see Agresti, 1992) is computationally expensive, whereas large-sample approximations such as those used by Blaheta and Johnson (2001) are potentially ill-suited for the rare counts that are the dominant problems in distributional similarity models.

Similarity between vectors Once the feature weights have been determined, we can use these vectors as a proxy for the meaning of a word (in a crude, impoverished sense of 'meaning'). In particular, we can use them to assess how similar two words are in their meaning, or apply clustering techniques to group words into semantically coherent categories.

The most popular vector similarity measure in Information Retrieval is the **cosine similarity**, which computes the dot product between two normalized vectors:

$$\mathrm{sim}_{\cos}(w_1, w_2) = \frac{\langle w_1, w_2 \rangle}{|w_1| \cdot |w_2|} = \frac{\langle w_1, w_2 \rangle}{\sqrt{\langle w_1, w_1 \rangle \langle w_2, w_2 \rangle}}$$

Since it explicitly normalizes the weight vectors, cosine probability can be applied to unnormalized vectors such as raw or logarithmic counts.

A second group of measures assumes that weights are roughly on a similar scale, which means that mutual information and similar statistics (which are similarly scaled for all words) can be used as-is, whereas raw counts or significance values would have to be scaled beforehand.

Lin's similarity measure Lin (1998a) counts the sum of weights for those features that are common to both target words, and divides it by the sum of weights for all features of each word:

$$\mathrm{sim}_{\mathrm{Lin}}(x_1, x_2) = \frac{\sum_{y \in F(x_1) \cap F(x_2)} w(x_1, y) + w(x_2, y)}{\sum_{y \in F(x_1)} w(x_1, y) + \sum_{y \in F(x_2)} w(x_2, y)}$$

where $F(x)$ designates all features that are associated with x (i.e., have a non-zero count or a positive mutual information value).

Closely related to Lin's measure is the **weighted Jaccard** measure (Grefenstette, 1992) that, instead of adding the weight values, takes minimum and maximum, respectively:

$$\text{sim}_{\text{WJaccard}} = \frac{\sum_y \min(w(x_1, y), w(x_2, y))}{\sum_y \max(w(x_1, y), w(x_2, y))}$$

Both Lin's similarity measure and Grefenstette's weighted Jaccard measure are unable to handle negative values as they can occur with (unbounded) mutual information in the case of dissociated attributes.

Other measures target the **divergence** of probability distributions, which are based on the notion that, when the actual distribution $p_x(y)$ of feature occurrences is replaced with a different distribution $p_{x'}(y)$, the information-theoretic coding becomes more inefficient the greater the difference between $p_x(y)$ and $p_{x'}(y)$ is.

The building block for the divergence measures is the *Kullback-Leibler* divergence between two probability distributions p and q:

$$D(p; q) = \sum_y p(y) \cdot \log \frac{p(y)}{q(y)}$$

The Kullback-Leibler divergence is a measure of dissimilarity: it assumes the smallest possible value, 0, when p and q are identical, and larger values the more dissimilar the distributions are. It is not a distance measure since it is not symmetric and does not fulfill the triangle inequality (see Manning and Schütze, 1999).

The Kullback-Leibler divergence, becomes infinite whenever $p(y)$ is non-zero for a feature where $q(y)$ is zero (which happens regularly with probability distributions that occur in target word - feature co-occurrences). Hence, two other measures based on, but distinct from, the Kullback-Leibler divergence have been proposed for the use in distributional similarity.

The first one is the *Jensen-Shannon divergence*, which measures the divergences to an averaged distribution:

$$\text{div}_{JS}(p_1, p_2) = D\left(p_1; \frac{p_1 + p_2}{2}\right) + D\left(p_2; \frac{p_1 + p_2}{2}\right)$$

Lee (1999, 2001) argues that it is preferable to approximate the actual Kullback-Leibler divergence more closely by skewing the averaged distibution towards the word one wants to compare it with:

$$\text{div}_{\text{skew}-\alpha}(p_1, p_2) = D(p_2; \alpha p_1 + (1 - \alpha) p_2)$$

Like the original Kullback-Leibler divergence, this measure is asymmetric, but in contrast to the former, it is always well-defined and has a finite value.

Weeds and Weir (2005) apply the familiar notions of precision and recall to distributional similarity in their framework of **co-occurrence retrieval**: In analogue

to the symmetrical Jaccard-like measures of Grefenstette and Lin, they define *additive* and *difference-weighted* measures for precision and recall:

$$\begin{aligned}
\mathrm{prec}_{add}(x_1, x_2) &= \frac{\sum_{y \in F(x_1) \cap F(x_2)} w(x_1, y)}{\sum_{y \in F(x_1)} w(x_1, y)} \\
\mathrm{recl}_{add}(x_1, x_2) &= \frac{\sum_{y \in F(x_1) \cap F(x_2)} w(x_2, y)}{\sum_{y \in F(x_2)} w(x_2, y)} \\
\mathrm{prec}_{dw}(x_1, x_2) &= \frac{\sum_{y} \min(w(x_1, y), w(x_2, y))}{\sum_{y} w(x_1, y)} \\
\mathrm{recl}_{dw}(x_1, x_2) &= \frac{\sum_{y} \min(w(x_1, y), w(x_2, y))}{\sum_{y} w(x_2, y)}
\end{aligned}$$

By combining precision and recall – using both a weighted arithmetic and the harmonic mean – Weeds and Weir are able to create a whole range of similarity measures (both symmetric and asymmetric) that put their emphasis on different aspects of the similarity between the respective weight vectors.

Dimensionality reduction and Re-estimation techniques Beyond the weighting techniques and similarity functions described above, some approaches make use of global properties of the vectors to compute a modified version of these vectors that are less noisy than beforehand.

One technique that is popular within the domain of window-based methods is **singular value decomposition** (SVD), in a group of techniques usually referred to as Latent Semantic Analysis (or Latent Semantic Indexing when applied to Information Retrieval).

Latent Semantic Indexing was born out of a desire to make search on large document collections computationally less costly (Dumais *et al.*, 1988; Deerwester *et al.*, 1990). Using singular value decomposition (SVD), it computes a low-rank approximation to the original document matrix.

Starting from a document matrix $A \in \mathbb{R}^{m \times n}$ (where $A_{i,j}$ could correspond to the count for word i in document j), SVD creates a decomposition $A = U^T \cdot S \cdot V$, where U and V are orthogonal matrices (i.e., $U^T = U^{-1}$ and thereby $UU^T = U^T U = I$) and S is a rectangular matrix where only diagonal entries are non-zero.

Setting to 0 all but the r largest entries of S's diagonal (where r is typically on the order of several hundred, whereas vocabulary and document collections have a size in the range of $10^4 - 10^5$) creates a low-rank approximation to A (which allows us to eliminate some rows from U and V, since the corresponding entries in the new matrix S' are all zero). With respect to all rank-r approximations, this matrix, $A' = U^T \cdot S' \cdot V$, is closest to the original matrix A.

The resulting changes from the full matrix A to the reduced-rank matrix A' can be seen as an approximation to *query expansion*, a useful technique in information

retrieval where the search query is enriched with synonyms or related words to words of the original query. Continuing our earlier example, if we posit a document collection which contains several documents about space travel, some of them containing the word *astronaut*, whereas others contain the word *cosmonaut*. A query for *astronaut* would not find any of the *cosmonaut* documents. Query expansion would add the synonym *cosmonaut* to the query, thereby increasing the recall of wanted documents. Using the query on the reduced-rank document-term matrix A' would result in the *cosmonaut* documents as well, since both kinds of space travel texts are otherwise very similar and the *astronaut/cosmonaut* distinction would only be found in the vectors corresponding to smaller singular values. Note that the difficulty for the document-term matrix to capture similarity between synonyms lies exactly in the fact that there is no generalization over contexts (which in this case are documents), and this is precisely why singular value decomposition makes a very important difference here.

Schütze (1992) also proposes a model that uses an SVD-based approximation of word cooccurrences (using a window-based approach) to perform word sense disambiguation by clustering occurrences of an ambiguous word by their context words and labeling those using the occurrences that are part of the training data. Schütze explicitly mentions that his use of SVD is not to overcome sparse data as it does in Dumais et al.'s Latent Semantic Indexing approach, but rather to ease the manipulation of the context vectors.

Geffet and Dagan (2004) compute new weight vectors for terms using a technique that corresponds to the approach of *blind relevance feedback* in information retrieval:[2] Starting from the assumption that the most highly weighted features are not necessarily informative, but that the most highly ranked semantic neighbours are good enough that these neighbours can be used to find more informative features.

Geffet and Dagan therefore replace the original weights in the vector by **relative feature focus** which uses features that the target word shares with closely related words:

$$w_{\text{rff}}(x, y) = \sum_{x' \in F^{-1}(y) \cap N(x)} sim(x, x')$$

where $F^{-1}(y)$ denotes the set of target words that have a positive weight for the feature y and $N(x)$ denotes the most closely related target words to a given target word x. For the set of most closely related words, Geffet and Dagan simply use all words that have a Lin similarity measure of at least 0.04.

Using an 18 million tokens subset of the Reuters corpus (Rose *et al.*, 2002), Geffet and Dagan found that RFF-weighted vector similarity was judged substan-

[2]Relevance feedback is a technique where the user selects relevant documents and the system extracts new query terms from these documents. Because this technique proved to be rather successful at improving search results, but user input is necessary for it, researchers came up with the idea of simply using a few most highly ranked documents which most likely would have a large fraction of relevant documents but would not require user input – *blind* relevance feedback.

tially better on a substitutability criterion than the using the baseline similarity measure based on mutual information weights (64% versus 54% substitutable terms among the 10 most similar neighbours returned by the similarity measure).

Comparative evaluation Within the general approach of distributional similarity, there is ample room for variation, in the inventory and acquisition method for the grammatical relations, in the weighting function used for the counts, as well as the similarity function used to derive similarity values from weight vectors. Several authors have compared at least some of the ingredients mentioned above in their reviews, and since it is not a goal of this thesis to try out all and every variant possible from the above ingredient, I will use this research to motivate the choice of a small number of reasonable measures.

Lin (1998a) used mutual information between word pairs in several relations output by his parser (Lin, 1998b), computing the similarity of the vectors of mutual information values with a variant of the weighted Jaccard measure described as $\mathrm{sim}_{\mathrm{Lin}}$ above and positive mutual information as the weighting method.

Lin compares his similarity metrics to several others, including the cosine measure and Hindle's original measure, and finds that his measure performs best (in terms of correlation coefficient) both in comparison to a WordNet-derived similarity measure and in comparison with one derived from the Roget thesaurus, even though (as Lin notes) the performance of his, Hindle's and the cosine similarity measures are relatively close.

Lin also found that the distributional similarity measures had a greater correlation with the WordNet-derived similarity measure than between WordNet- and Roget-derived similarity measure, whereas the similarity with the Roget-derived similarity measure was less than that value.

Curran and Moens (2002a) use a set of 70 noun terms balanced for frequency, number of senses, depth in the WordNet hierarchy and position within WordNet. Using an evaluation against three electronic thesauri (Macquarie, Roget's thesaurus, and the Moby thesaurus), they find that a combination of t-test weighting together with the weighted Jaccard metric yields the best results among the possibilities for weighting and similarity measures that they tried, with mutual information and Lin's Jaccard variant (i.e., the components used in Lin's 1998 similarity measure) relatively close.

The most comprehensive study comparing distributional similarity measures in a common framework to date is the work of Weeds (2003). Weeds uses the analogy of retrieving weighted sets of items (in this discussion: the collocates), and comparing the weighted set of collocates retrieved for a candidate word with that of the target word.

In this framework, which she calls *co-occurrence retrieval*, parametrization yields a large, varied space of models which either subsumes existing models such as Lin's or can be tuned to approximate them. Weeds starts from simple figures-of-merit for comparing the co-occurrences of one item with another item (in terms

of precision and recall values), which can then be combined in a flexible way.

Weeds then uses a weighted mean of Precision, Recall, and F-measure (as the unweighted harmonic mean of Precision and Recall) to get a continuum of measures that have different properties. Weeds then tests the different measures in several settings:

- *Similarity-based smoothing of language models*, in which the distribution of verb-object pairs is smoothed with by interpolating the actual distribution of verbs seen with a noun with the distribution observed with semantically close neighbours.
- *Pseudo-Disambiguation experiments*, in which a noun-verb tuple (n, v_1) is taken and a second verb v_2 with a similar frequency to v_1 is chosen randomly; using the nearest neighbours of n, the model then has to predict which of v_1, v_2 was the verb that was actually seen with n.
- *Correlation with Human Similarity Judgements* using the data set of Miller and Charles (1991)
- *Comparison of neighbour sets* between the 200-most-similar words according to the wordnet-based similarity measure of Lin (1998c) and the 200-most-similar items from the distributional model
- Retrieving WordNet hypernyms and hyponyms.
- Detecting real-word spelling errors (i.e., a misspelled word resulting in another dictionary word), similar to Budanitsky and Hirst (2001).

Although tuning a measure inside her model of Co-Occurrence Retrieval usually yields the best achieved results, Lin's measure performs quite well on all tasks, with a performance relatively close to the best attainable using her model. This could be seen as an indication for the general usefulness of Lin's approach as a general-purpose distributional similarity measure.

Versley (2008a) uses the UK-WaC corpus and the Google 1T n-gram dataset to compare relation-based approaches using pattern-based extraction of syntactic or semantic relations, on the task of clustering a set of concrete nouns and verbs. Even though the simpler window-based approach could be used on the much larger n-gram data, the window-based approach on the n-gram dataset does not exceed the result for the best-performing single relation on the DE-WaC corpus, even when using a variant of singular value decomposition that downweighs the predominance of few topic distinctions by rescaling the diagonal vectors.

For the distributional similarity measures discussed in section 6.2, I decided to implement something relatively close to Lin's similarity measure – using grammatical relations, as well as mutual information weighting and Lin's Jaccard variant – the main reason being that, at least according to the study by Weeds (2003), there

is considerable variance among different tasks as to which distributional similarity measure works best, and Lin's measure seems to work near-optimally on all tasks. Additionally, Padó and Lapata's relation-free measure based on syntactic neighbourhoods was chosen as a second distributional similarity measure as it seems to be the best option for a relation-free measure.

6.1.2 Pattern-based approaches

Another approach to approximate semantic knowledge from raw text has been introduced by Hearst (1992), and popularized for the use in coreference resolution by Markert and Nissim (2005). Hearst noted that patterns like the ones used for extracting relations encoded in dictionary definitions could also be used to extract semantic relations from free text, provided that they:

- occur frequently and in many text genres
- (almost) always indicate the relation of interest and
- can be recognized with little pre-encoded knowledge.

Hearst notices that such patterns can be found by gathering a list of terms for which the desired relation (most commonly, hyponymy) holds, possibly using existing patterns to find candidate examples, and then finding contexts in the corpus where these expressions occur syntactically near one another; in a further step, the environments of these co-occurrences are recorded and one tries to find the commonalities among these environments, assuming that common patterns can be a good indicator for the relation of interest.

Hearst proposes five such patterns, which are referred to in the literature as the *Hearst patterns*:

- `NP` such as (`NP`,)* (and|or) $\texttt{NP}_n$
- `NP`, (`NP`,)* or other `NP`
- `NP`, (`NP`,)* and other `NP`
- `NP`, including (`NP`,)* (and|or) $\texttt{NP}_n$
- `NP`, especially (`NP`,)* (and|or) $\texttt{NP}_n$

Berland and Charniak (1999) use Hearst-like patterns to find instances of part-of relations using several patterns indicating an associative relation (*whole*'s *part*, *whole-NP* of (the|a) *part*, *part* in (the|a) (JJ|NN)* *whole*, *part-pl* of *whole-pl*, *part-pl* in *whole-pl*).

Using a large (100 million words) newspaper corpus, they used association statistics to find nouns that denote parts of one of six target concepts (*book*, *building*, *car*, *hospital*, *plant*, and *school*), using both log-likelihood and a lower estimate of mutual information (realized as the lower bound of a confidence interval;

they call this method "significant-difference", or sigdiff). Berland and Charniak used an additional filter to remove all quality-denoting words based on suffixes such as *-ing*, *-ness* and *-ity*, and found the whole approach to work rather well (55% precision compared to human judgements).

Poesio *et al.* (2002) use patterns indicating part-hood (the *part* of *whole*, *part* of *whole*, *whole*'s *part*) to resolve meronymic associative bridging cases, using mutual information to identify the most strongly associated antecedent candidate, and report that this approach outperforms not only WordNet (which has insufficient coverage for meronymic relations), but also an approach based on a HAL-like association measure.

Markert and collaborators (Markert *et al.* 2003, Markert and Nissim 2005) note that the strong semantic relation that is implicitly expressed in the case of *other*-anaphora (i.e., finding what the contrast set for an NP like *other risk factors* may be) or coreference, or associative bridging, is likely to be structurally expressed. Given enough text, it is not only possible to use this information directly, but also desirable, since context-dependent assertions such as *age* being a *risk factor* in the following context:

(6.3) You either believe Seymour can do it again or you don't. Beside *the designer's age*, **other risk factors for Mr. Cray's company** include the Cray-3's [...] chip technology.

Markert *et al.* also compare using the Web with results that can be achieved using the British National Corpus (Burnard, 1995), a large text corpus of about 100 million words. While Markert *et al.* (2003) found that using the Web did not perform better than Poesio et al.'s (2002) pattern-based search on the BNC for associative bridging, Markert and Nissim (2005) found that the web-based approach did have much better recall for hyponymy relations, even when allowing for some flexibility in the BNC matcher by allowing modifiers in between.

Several approaches use Hearst- or Hearst-like patterns to construct or populate ontologies: Evans (2003) uses an algorithm that extracts hypernym candidates for named entities in a text using web queries for occurrences of a Hearst pattern with the hyponym position filled by the named entity in question, and then proceeds by clustering the named entities according to their possible hypernyms using agglomerative average-link clustering. The resulting clusters could then be labeled using the most specific concept from WordNet that covers all hypernyms in the cluster.

Caraballo (1999), on the other hand, uses distributional information about coordination to perform a variant of agglomerative average-link clustering and then uses Hearst patterns to find labels for the nodes: she collects possible hypernyms for single words using the "X_1, X_2, and other Ys" pattern (which is applied to parsed trees), and uses the three hypernyms that are most important in terms of word types. She also compresses the binary tree that is the result of the agglomerative clustering procedure by merging nodes that have the same hypernym labels.

Cimiano and Staab (2004) present an approach that uses hyponymy-indicating patterns to populate an ontology with named entities: they sum up match counts

returned by the Google API over all patterns used (filling pre-defined query templates with the entity to classify and a concept of the ontology), and keep the those of the m highest-ranked concepts that have a count sum over a certain threshold θ. Classifying entities into a set of 59 concepts, they reached an accuracy of 24.9%.

In addition to Hearst's original set of patterns, Cimiano and Staab also use appositional and premodifying patterns (the Y X, the X Y) for names, as well as appositional (X, a Y) and copula constructions (Y is a X).

Cederberg and Widdows (2003) use an LSA-based similarity measure to rank possible hypernyms for noun patterns found in a corpus; they then use these hypernyms to rank possible hypernyms, and use these as candidate labels for clusters based on similar coordinated nouns. In a second step, they use the label with the maximum similarity to the other words of the cluster.

Several authors propose methods to acquire more patterns by somehow automating Hearst's proposed method to bootstrap patterns and instances:

Geleijnse and Korst (2006) present an algorithm to bootstrap both patterns indicative of a relation and instances of the relation (i.e., word pairs between which the relation holds), assuming that the sets I_q and I_a from which the instances on the left/right side of the relationed pair come are known.

Geleijnse and Korst start by querying a search engine for occurrences of one term near the other, using the queries `allintext:"`$X \star Y$`"` and `allintext:"`$Y \star X$`"` (where the `allintext:` operator exclude matches from the title of a web page) for a given pair of terms that are known to be related and retrieve the documents returned for this search; they then extract all the phrases matching the queried expression, removing all matches that cross a sentence boundary.

In a second step, they score the highest-frequency patterns according to a combination of measures for frequency (returning a high number of relevant matches), high precision (i.e., only returning relevant-looking matches), as well as spread (returning as many distinct relation instances as possible):

- their frequency criterium is just the raw count of instances (capped to 1000 per queried pair because the Google API does not return more matches).
- their precision criterium is a micro-average of the proportion of matches returned for each member of a (pre-specified) subset I'_q of I_q that belong to the set I_a of related candidates.
- the spread is measured as the sum over number of distinct relation instances of I_a that were found by querying the pattern with the instances from I'_q.

Gelijnse and Korst find that using the pattern bootstrapping with the term "country" and a list of countries yields a number of reasonable patterns indicating hyponymy, remarking that patterns of the form "is a *adjective* country" occurred relatively frequently, which might be exploited by additionally using a part-of-speech tagger to yield more abstract patterns. In a second experiment, they

investigated bootstrapping patterns for a *restaurant-country* relation (roughly corresponding to the TREC 2004 question "What countries is Burger King located in?"), and found that several kinds of restaurants were found (among non-targeted terms such as names of cuisines, or city names), and that the extracted list had good precision and recall for the countries related to 'Burger King'.

Snow *et al.* (2005) train a supervised classifier based on a large dataset using pairs of related unambiguous nouns extracted from WordNet as positive examples, and random pairs of nouns as negative examples. The random pairs are sampled to match the 1:50 ratio between related and non-related pairs encountered in words randomly chosen from the same paragraph. Using 6 million sentences of newswire text parsed with MiniPar, they represent each pair of nouns by the dependency paths between them found in the parses where both nouns occur in the same sentence and are no more than four dependency links apart. Additionally, they extend the paths by so-called satellite links, additional dependency links that are connected to either noun and allow to cover the syntactic patterns for "X *and other* Y" (where *other*,A:mod:N is the satellite link, or *Such* Y *as* X, where *such*,PreDet:pre:N is the satellite link).

Testing the resulting logistic regression classifier both by cross-validation on the WordNet-based pairs and on a test set of hand-annotated pairs, they find that their model is able to effectively combine the occurrence information of a large number of patterns (about 70.000 dependency paths), yielding better precision and recall than either single patterns alone or the set of patterns originally proposed by Hearst.

Snow *et al.* are able to additionally improve the recall of their classifier by using a similarity-based smoothing scheme including likely coordinate sisters of the (candidate) hyponym:

$$P_{\text{new}}(n_i \overset{H}{<} n_k) := \lambda_1 \cdot P_{\text{old}}(n_i \overset{H}{<} n_k) + \lambda_2 \cdot \sum_{n_j} P_{\text{sim}}(n_j|n_i) P_{\text{old}}(n_j \overset{H}{<} n_k)$$

where $P_{\text{sim}}(n_j|n_i)$ is a distribution over the most similar words of n_i that is determined by a distributional similarity measure similar to Lin's (1998a).

While both the approaches of Markert and Nissim (shallow pattern search on the World Wide Web) and of Snow *et al.* look very promising, it is important to note that their approaches are in a sense complementary, but not easily combinable: while Markert and Nissim's approach of using a search engine's efficient indexing infrastructure is difficult to scale to the number of patterns that Snow *et al.* use, Snow et al.'s use of full parsing limits the size of the data that can be processed.

In the case of distributional similarity measures based on grammatical relations, approximate parsing methods are used that provide a slightly noisier version of the grammatical relations that would be extracted by a full parser while being computationally inexpensive, such as Curran's reimplementation of Grefenstette's Sextant parser (Curran and Moens, 2002b), a parser that performs well enough on (nearly) all grammatical relations is much harder to achieve.

6.1.3 Cross-sentence associations

While looking at the (syntactic) neighbourhood of a word will give a good representation of how such a word is used, the bridging relations we are interested in, also usually signal a meaningful semantic relation.

Bunescu (2003) uses cross-sentence associations in patterns to detect associated pairs via web pattern search. Bunescu uses two patterns which are expected to be found much more frequently with associated pairs:

- X. The Y *verb*
- X NEAR the Y

Using pointwise mutual information, Bunescu's approach yields two association statistics between the two nouns. Bunescu then evaluates his approach using 686 modifierless noun phrases of which 324 were classified as anaphoric (i.e., which had a preceding trigger noun in the list of preceding noun phrases), whereas the others had no triggering noun phrase either because the trigger was a verb, a whole phrase or the general situation described, because the noun phrase was uniquely referring (as in *the moon*), or because the noun phrase was non-referring (e.g., occurring inside an idiomatic phrase). In the evaluation, each of the 686 potentially anaphoric noun phrases is assigned the most closely associated of the 50 preceding noun phrases as textual antecedent, filtering out cases where the association score was below a threshold.

Bunescu finds that the phrase-based approach works significantly better, with a precision of 53% at a recall of about 22%, than either the approach using the NEAR operator or the approach of Poesio *et al.* (1998).

Garera and Yarowsky (2006) use an approach very similar to Bunescu's to model hyperonymy relations as they would be used for coreferent bridging. Using the LDC Gigaword corpus[3] as the textual base they computed pointwise mutual information scores between pairs of anaphors and potential non-same-head antecedents. This was achieved by collecting pairs of noun phrases where one was a definite without a same-head antecedent, on one hand, and the other one was a noun phrase occurring in a window of the last two sentences.

As an example, let us consider the following example from Garera and Yarowsky (taken from the LDC Gigaword corpus):

(6.4) . . . *pseudoephedrine* is found in an allergy treatment, which was given to Wilson by a doctor when he attended Blinn junior college in Houston. In an unanimous vote, the Norwedian[sic] sports confederation ruled that Wilson had not taken **the drug** to enhance his performance.

In the example, *the drug* would have the potential antecedents *pseudoephedrine, allergy, Blinn, college, Houston, vote, confederation, Wilson*, and thus this example sentence would yield, among others, pairs like *pseudoephedrine-drug*, *Houston-drug* and *allergy-drug*. All these candidate pairs would then be collected, and the

[3] David Graff (2003): English Gigaword, LDC2003T05

mutual information statistic for the word pairs over the whole corpus would be calculated, in the hope that *pseudoephedrine-drug* would get a higher score than other items.

Garera and Yarowsky evaluate their approach on 177 examples of an anaphoric definite NP where the antecedent was in one of the two preceding sentences. Comparing it to a method using hyperonymy relations in WordNet, as well as Markert and Nissim's pattern-based approach (using the Gigaword corpus for pattern search, instead of using the World Wide Web), they find that a combination of WordNet and their TheY model works better than either WordNet alone or a combination of the pattern-based approach and WordNet. Of the single information sources, WordNet performs best, followed by their TheY model and the pattern-based approach, which is consistent with Markert and Nissim's observation that the pattern-based approach fares less well when used on a large corpus than when used on the World Wide Web.

6.2 Corpus-based Semantic Features for German

Unsupervised learning methods – in particular, distributional similarity and association measures such as the ones that are implemented in this section – are instrumental to improving the recall of antecedent selection since their coverage only depends on the unlabeled data that is used (as opposed to manual labor that is necessary for hand-crafted resources). In contrast to hand-crafted resources, the similarity relations that are within reach of distributional similarity are not constrained by taxonomical organization principles and may be better suited for context-dependent relations such as the near-synonymy between *Album* and *CD*.

Among the distributional approaches, I will explore the hypothesis that relation-based approaches, which are more popular in the computational linguistics community, are more useful than relation-free approaches, which are more popular among the cognitive psychology community, by evaluating both a relation-based similarity measure inspired by Lin (1998a), which is among the best-performing measures for all the tasks investigated by Weeds (2003), and a syntax-based relation-free measure such as the one presented by Padó and Lapata (2003) in the context of antecedent selection in a later section of this chapter 6.4, but also by a qualitative inspection of the most-similar items retrieved in the discussion at the end of this section.

Several factors contribute to the quality of the similarity measure: one is the use of fast, accurate parsing techniques to extract the grammatical relations that are used as features, which is described in subsections 6.2.1 and 6.2.2; the other is the use of compound splitting to mitigate the effect of German synthetic compounds on vocabulary growth (see subsection 6.2.3).

As a prerequisite of relation-based distributional similarity measures such as the one of Lin (1998a), but also for models of selectional preferences (such as the one of Rooth *et al.*, 1999), it is necessary to process corpus data and extract syntagmatic relations (in our case, grammatical relations).

Curran, who conducted a large-scale study of syntactic preprocessing techniques for English similarity measures (Curran and Moens, 2002b; Curran, 2004), notes that different kinds of tradeoff are involved here: usually, parsers that are more accurate take longer time, while the data they provide is also less noisy; In this manner, the process of extracting grammatical relations is bounded both by the available quantity of high-quality text (as crawling the web, or other means of acquiring additional texts, would also imply a significantly lower text quality), but also by the time available for parsing the texts.

What is necessary to implement similar techniques for German that result in usable distributional similarity measures to be incorporated in a coreference resolution system? One aspect of this is to be found in essential differences between German and English, which need to be taken into account – essentially, richer morphology, which makes lemmatization an absolute necessity, and also the presence of synthetic compounds, which has the potential for creating a sparse data problem. The other aspect of this is that essential infrastructure which readily exists for English to the point where one can just select among several available off-the-shelf components, whereas their German counterpart is either non-existent, performs significantly worse, or is not available to the research community at large due to usage restrictions.

6.2.1 Parsing the taz Corpus

As a source of grammatical relations for distributional similarity measures, the parser would have had to fulfill several criteria:

- have good accuracy on the targeted grammatical relation in typical sentences
- make it possible to parse a large quantity of text (i.e., the 200 million words from the taz corpus, or a substantial part thereof) in a reasonable time.

  For comparison: a parser that takes about 1 second per sentence will parse 11 million sentences in about 127 days, which, taking into account parallelization, is a realistic amount. Before the integration of parse predictors into CDG, a realistic expectancy of time would have been 60 seconds per sentence (Kilian Foth, p.c.), after their integration (cf. Foth and Menzel 2006) this was down to 10 seconds, which would still have been a disproportionate amount of time for the large-scale parsing task I had in mind.

As of 2005, shift-reduce dependency parsers were not the first choice for such parsing tasks, as the accuracy for case identification in them (and as a result, of the grammatical functions we are interested in) was not sufficient. Furthermore, the available interpretations were either rather slow (in the case of Nivre's original implementation) or were relatively inaccurate (in the case of Foth's implementation based on count statistics with simple backoff).

Incremental shallow parsing, such as the approach of Müller (2004a), which later has been used by Wunsch and Hinrichs (2006), could have been a compelling

alternative, but Müller's implementation is largely underdocumented and more of a 'prototype' (Frank Müller, p.c.[4]) which would be nontrivial to set up.

A good compromise between accuracy and speed seemed to be the strand of research started by Klein and Manning (2003a), who use unlexicalized PCFG parsing with a grammar that contains more information than the original treebank, and which can be obtained by augmenting the original symbols with additional linguistically motivated information. The parser by Schiehlen (2004a) (and later, the one by Dubey 2005) are a good example that the approach of PCFG parsing based on a transformed treebank could work well for German.

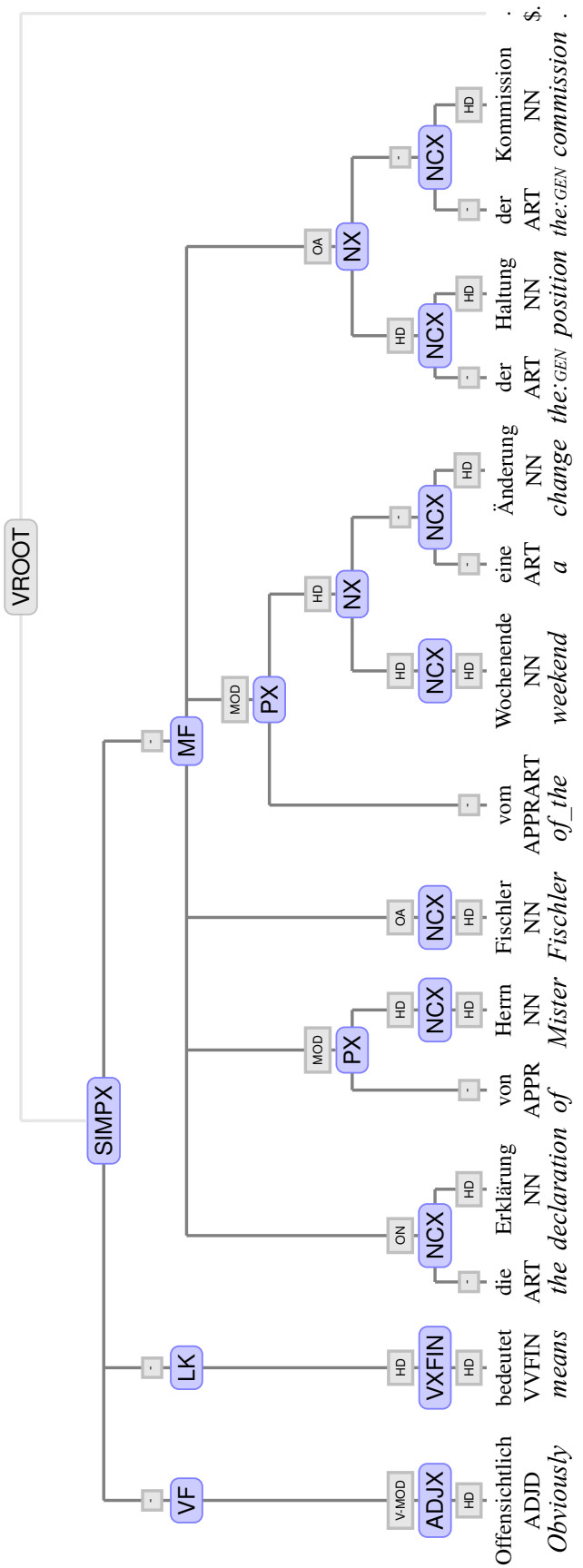
Like Schiehlen, I decided to use BitPar (Schmid, 2004), an unlexicalized PCFG parsing engine that was freely available in source form and is well-suited for practical parsing by allowing arbitrarily large sets of symbols and achieves reasonable efficiency by first computing a non-probabilistic forest through CFG parsing, which is represented as bit vectors, performing top-down filtering to eliminate non-productive chart edges which cannot lead to a complete parse, and subsequently constructing a probabilistic parse forest from the filtered forest.

In comparison, LoPar, another parser that is freely available, only allows a limited length of symbols, and consumes an order of magnitude more memory because its left-corner filtering is less effective than the top-down filtering that BitPar uses.

Modifications to the engine of BitPar were limited to measures towards improving the treatment of unknown words, by means of implementing an improved unknown word model (see pages 177ff.) and filtering of possible tags by means of a morphological analyzer (see page 178). Limiting the scope of modifications to the parsing engine is necessary, as such modifications are potentially costly in terms of development and debugging time. At the same time, the default unknown word model of BitPar is clearly inadequate, having only two distributions for upper- and lowercase words. (Incidentally, Schmid (2006) also adds a more adequate mechanism for unknown word prediction).

Because BitPar includes an efficient procedure for grammar binarization (cf. Schmid 2004), and the more hierarchical structure of the TüBa-D/Z is less prone to sparse data problems, use of binarization is limited to markovization in the case of coordinations, which behave very regularly yet yield arbitrary-length rules, and to account for linguistic regularities in the case of nonfinite verb clusters in the verbal complex.

[4] In an email from August 2005.



Obviously, Mr. Fischler's declaration on the weekend indicates a change in the views of the Commission.

Figure 6.2: Problems with the unmodified treebank grammar (vanilla)

Looking at parses from the unmodified treebank grammar induced from TüBa-D/Z (cf. figure 6.2), the most important problems seemed to be the following:

- With the richer morphology and **flexible word order** of German, grammatical functions (such as subject and accusative object) are no longer definable purely in terms of configuration (as in English, where a noun phrase as a verb argument on the left of the verb must be the subject). Instead, grammatical function is determined by **case**, which is *morphologically encoded*, and it is only in case-ambiguous clauses that argument order and selectional preferences play a role.

  Just adding edge labels to the nodes in the case of the unmodified treebank grammar leaves the parsing model without any indication of the morphology. In figure 6.2, only the subject is annotated correctly. "*Fischler*", which is an apposition to "*Herrn*", is annotated as the accusative object, while the actual accusative object "*eine Änderung ...*" is annotated as genitive postmodifier to "*Wochenende*", which would be highly implausible with morphological information.

- In the case of relative clauses, we get a **C field** that contains a (relative) noun phrase. The result of this was that in some parses, an ordinary noun phrase would be forming a C field in a nonsensical sentence structure.

- The VXFIN and VXINF chunks hide away the useful information from the verbs that would be helpful to identify general clause structure (i.e., *haben* and *sein* are usually auxiliaries to past participles, whereas *zu*-infinitives usually occur in subordinate clauses). Similarly, the LK and VC nodes hide information about the verbs which can help in the identification of clause boundaries in more complex cases, as well as for the identification of verb valence when additional information is added.

  Inconsistencies between the contents of LK and VC fields and clause structure are most often a problem in conjunction with unknown words being mis-analyzed as verbs, or (mediated by the lack of distinctions between different types of partial-clause conjuncts) in the case of coordination:[5]

  (6.5) a. ... [$_{\text{SIMPX}}$ die [$_{\text{FKONJ}}$ vom Museumspädagogischen
  ... that/REL by the Museumspädagogischer
  Dienst Berlin [$_{\text{VC}}$ freitags/VVPP]] bis [$_{\text{FKONJ}}$ sonntags
  Dienst Berlin Friday:GEN to Saturday:GEN
  angeboten wird]]
  offered PASS:3SG
  ... that is offered by the Museumspädagogischer Dienst Berlin on Fridays to Saturdays

[5] In the setting examined by Versley (2005), parsing of out-of-domain text without help from a morphological analyzer, such mis-analyses of unknown word are much more frequent than in an in-domain setting.

b. Im Westen [$_{\text{FKONJ}}$ soll der Eisenbahnring die A 20]
in the West shall:3SG the railway ring the A 20
und [$_{\text{FKONJ}}$ die A 26 überflüssig machen].
and the A 26 superfluous make.
In the West, the railway ring is to render superfluous the A20 and the A26.

- Certain noun phrase types such as parts of extraposed NP coordination ("*Er hat ihr eine Blume geschenkt [NX und Pralinen]*") or comparative clauses ("*Er ist größer geworden [NX als sein Bruder]*") have a very different behavior from normal noun phrases yet receive the same label. As a result, mis-parses frequently include such phrases in positions where they do not make sense.

A first group of transformations pertains to a more accurate reconstruction of noun phrase case and other properties: First of all, noun phrases are annotated with their case. Comparative or extraposed-conjunction noun phrases receive a distinct label (KomX in the case of comparative noun phrases and NXK in the case of extraposed NP coordination). Interrogative and relative pronouns, which can only occur in complementizer position (of questions and relative clauses, respectively) are also assigned distinct labels: Noun phrases containing an interrogative pronouns (what, which car, of which car) are labeled as NX-W, and relative pronouns and noun phrases with attributive relatives such as *dessen* and *deren*[6] are labeled NX-REL.

Noun Phrase Morphology As noted above, case information has to be derived from the morphological properties of the elements of the noun phrase. The PCFG grammars of Schiehlen (2004a) and Dubey (2005) add very little information to the treebank trees – Dubey marks the NP case on the determiner whereas Schiehlen marks case on the head noun. In contrast, both Müller (2004a) and Trushkina and Hinrichs (2004) use full morphological information to deduce the case of noun phrases, but use more elaborate lexical information. As the parser should run well on unrestricted text, using a backup lexicon with prefabricated morphological analyses for all words in the dataset, as done by Trushkina, as well as by Schulte im Walde (2000), would not be a good solution.

A good lexicon for the PCFG grammar should incorporate enough distinctions on the morphological level that there is sufficient information to model the case of noun phrases, but on the other hand, avoid the data sparsity issues that occur when the tagset grows too big. Avoiding data sparsity is important since an unseen $(word, tag)$ pair will lead to a parse that either uses the wrong POS tag for the word (and has a high chance of being wrong altogether) or no parse at all. It is therefore very important to keep the set of possible tags for one particular word

[6]The attributive relative pronouns *dessen* and *deren* would be translated as *whose* or *of which*, such as in *This is the man whose name I always forget.*

as small as possible if one wants to acquire a reasonably good lexicon from the treebank, or more generally, it is advisable keep the tagset as a whole small.

While there are three definite and five indefinite articles, five inflectional variants of adjectives and usually two or three inflectional variants for a given noun, the syncretisms of these forms usually do not coincide (cf. Trushkina 2004). As a result, a completely accurate (P)CFG grammar would need to distinguish case, number, gender and adjectival weak/strong inflection features, yielding 48 different possibilities. Partially underspecifying the case of nouns, as done by Schulte im Walde (2000), who uses a morphological analyser to find all the possible cases for a particular noun and then uses an underspecified tag, leads to a further explosion of the possible rules, which is undesirable.

In a first approach, determiners were annotated with case, number, and gender, whereas nouns would be annotated with their number and gender only. This excludes some erroneous analyses (such as an analysis of "der Abend" as genitive feminine), but sometimes gave unsatisfactory results in the case of unknown or ambiguous nouns. As a further improvement, inflected adjectives are annotated with their ending, which leads to a better analysis even for these cases.

Improving argument assignment To improve the assignment of verb arguments in the grammar, I wanted to use verb subcategorization information from the WCDG lexicon, so that accusative and dative arguments would be adequately recognized (consider the earlier example (5.9), where the distinction between apposition or two separate noun phrases is due to the arguments required by the verb). Looking at the TüBa-D/Z from the PCFG learning viewpoint, we can make out the following desiderata:

- The topological fields VF and MF, which contain the arguments of the sentence, do not pass this information on to the level of the sentence node. The result of this is that even with case marking, we would still have sentences with two subjects or direct objects.
- In order to prevent sparse data problems, it would be desirable to have a small number of verbal categories, and some abstraction akin to verb chunks; on the other hand, it is also necessary to provide enough context to account for verb valence and valence-changing operations such as passives.

By **annotating topological fields** with the arguments that occur in them (i.e. $\mathrm{MF}_{\langle \mathrm{OA,OD} \rangle}$ for a middle field containing both an accusative and a dative object), it is possible to exploit the regularities modeled by the topological fields model while at the same time preventing information loss that prevents the original treebank grammar from excluding nonsensical argument structures.

To account for the regularities in the **left and right sentence brackets**, while also excluding overgeneralizations, it is useful to remember that a clause may contain multiple non-finite verbs, but only one finite verb, which may either be part of the left or the right sentence bracket:

(6.6) a. daß er das Examen [$_{\mathrm{VC}}$ bestehen können wird]
that he the exam pass be able will:3SG
that he will be able to pass the exam.
b. Er [$_{\mathrm{LK}}$ wird] das Examen [$_{\mathrm{VC}}$ bestehen können].
he will:3SG the exam pass be able
He will be able to pass the exam.

It is possible to deduce the verb valencies from the structure of the verbal chunk, as realized in the HPSG account of Hinrichs and Nakazawa (1989). However, for regularity reasons, we need to exclude multi-verb chunks in the left sentence bracket:

(6.7) *Er [$_{\mathrm{VX}}$ bestehen können wird] das Examen.

For this reason, I merge only non-finite verbs into verb chunks (earlier examples repeated), which are annotated with the type of the verb chunk, similar to the STTS subcategorization of verbs into full/modal/auxiliary verbs on one hand, and finite/infinitive/*zu*-infinitive/perfect forms on the other hand:

(6.8) a. daß er das Examen [$_{\mathrm{VXVI}}$ bestehen können] [$_{\mathrm{VXAF}}$ wird]
b. Er [$_{\mathrm{VXAF}}$ wird] das Examen [$_{\mathrm{VXVI}}$ bestehen können].
c. *Er [$_{\mathrm{VXVI}}$ bestehen können] [$_{\mathrm{VXAF}}$ wird] das Examen.

The same binary-branching structure that allows us to summarize multiple verbs (bestanden$_{\mathrm{VXVP}}$·haben$_{\mathrm{VXAI}}$ → bestanden haben$_{\mathrm{VXVI}}$), also allows us to propagate valence information in much the same way:

$$\text{bestanden}_{\mathrm{VXVP:a}} \cdot \text{haben}_{\mathrm{VXAI:haben}} \rightarrow \text{bestanden haben}_{\mathrm{VXVI:a}}$$

In the example, the accusative object valence (:a) is percolated from the main verb to the verb group; In contrast, a dynamic passive with *werden* would remove the accusative valence for the verb group.

The subcategorization of verbs also allows to account for valency changing operations in passives, most importantly the progressive passive:

$$\text{gegeben}_{\mathrm{VXVP:ad}} \cdot \text{worden}_{\mathrm{VXAP:werden}} \rightarrow \text{gegeben worden}_{\mathrm{VXVP:d}}$$

To subcategorize verbs by their valency, it would be possible to annotate each verb token with the arguments it has in the current verb, which would be problematic in the case of coordination (where it would sometimes not be clear if arguments are shared across the conjuncts or not), and it would also create a problem of data sparsity especially for verbs that multiple possible valency frames.

The WCDG parser uses one single tag for each verb that approximates the possible valencies that it can have – for example, `ac` would be used for a verb that requires either a direct (accusative) or clausal object, whereas `a+d?` would be used for a verb that requires an accusative object and allows a dative object.

WCDG valencies consist of a string of letters, namely for accusative object (`a`), dative object (`d`), *dass*-clausal object (`c`), infinitive clause (`i`), V2 clause (`s`), or prepositional objects (`p`, with the filler of `objp` property as the preposition). Arguments can be marked as obligatorily reflexive by prefixing them with `r` (yielding

`ra` for a reflexive accusative), they may be marked as optional by suffixing with a `?`, and several of them can be used to express an alternative (for example, *leugnen*/deny has the tag `aci`, meaning that it takes either an accusative, a *dass*-clause or an infinitive clause as its argument). Conjunction of valencies may be expressed with the + operator, as in `a+d` for ditransitive verbs.

Experiments with the full WCDG tags, which include information about nominal, prepositional and clausal arguments requirements, showed that the full tags would create a massive sparse data problem. After some experimentation, this was reduced to a much smaller tag set which indicates whether the verb in question can have accusative and/or dative objects (i.e., just ε, `a`, `d` and `ad`).

Refining Sentential Categories The TüBa-D/Z uses several categories for clauses or conjoined parts of clauses:

- SIMPX for sentences and most subclauses
- R-SIMPX for relative clauses
- FKOORD for a partial clause coordination, where the single conjuncts are either FKONJ nodes (if they contain material from multiple topological fields) or simply MF or VF nodes if they contain material from a single topological field.
- FKONJ nodes contain multiple topological fields which are part of one single conjunct.

The way that FKOORD and FKONJ are defined, they can contain quite a variety of material, including verb-first and verb-last (partial) clauses or VPs. Complex sentences often show a combination of clausal coordination, clausal subordination, and noun phrase coordination, and it is especially important to enforce the strong linguistic constraints on clausal structure in order to avoid misparses.

To improve the parsing quality for conjunctions, FKONJ, SIMPX and FKOORD are subcategorized according to the material they contain:[7]

- finite verb-first (partial) clause (v1)

  (6.9) Der Verkehr [$_{\text{FKOORD:v1}}$ [$_{\text{FKONJ:v1}}$ verursacht nahezu die Hälfte
  the traffic cause:3SG nearly the half
  der Energiekosten] und [$_{\text{FKONJ:v1}}$ verbraucht nahezu die
  the:GEN energy costs and consume:3SG nearly the
  Hälfte des Erdöls]].
  half the:GEN crude oil.
  Transport takes almost half the energy cost and it uases almost half the oil.

[7] The following examples come from the EuroParl corpus.

- finite verb-last (partial) clause (v-fin)

  (6.10) Ich kann es jedoch nicht ertragen, wenn [$_{\text{FKOORD:v-fin}}$
  I can:1SG it however not tolerate when
  [$_{\text{FKONJ:v-fin}}$ man einen unabhängigen Ethik-Ausschuß fordert] und
  one an independent ethics committee demand and
  [$_{\text{FKONJ:v-fin}}$ dann seine Erkenntnisse nicht akzeptiert]].
  then its findings not accept:3SG.
  I cannot tolerate it, however, when people demand an ethics committee and then refuse to accept its findings.

- infinitive (infinitive)

  (6.11) Sie wollen [$_{\text{FKOORD:infinitive}}$ [$_{\text{FKONJ:infinitive}}$ mit den Hunden
  They want:3PL with the dogs
  jagen] und [$_{\text{FKONJ:infinitive}}$ mit den Hasen rennen]].
  hunt and with the hares run.
  They want to hunt with the hounds and to run with the hare.

- past participle (past-participle)

  (6.12) Japan und die Vereinigten Staaten haben [$_{\text{FKOORD:past-participle}}$
  Japan and the United States have:3PL
  [$_{\text{FKONJ:past-participle}}$ Vorschriften zu diesen orphan drugs genannten
  regulations about these orphan drugs called
  Medikamenten erlassen] und [$_{\text{FKONJ:past-participle}}$ ihnen einen
  pharmaceuticals enacted and them a
  besonderen Status eingeräumt].
  special status awarded.
  Japan and the United States have enacted regulations for these pharmaceuticals, called 'orphan drugs', and awarded them a special status.

- for coordination of single fields, FKOORD is annotated with the grammatical functions of the arguments occurring in that field.

Unknown words A general problem in parsing which is often overlooked is the treatment of unknown words, for which the parser must guess the syntactic category. Previously successful approaches for this use part-of-speech tagging, as in the case of Dubey and Keller (2003), where the part-of-speech tagger usually uses suffix information of the word to influence the decisions made, constructing a dictionary using a morphological analyser, as done by Schiehlen (2004a), Trushkina (2004), or Schulte im Walde (2000), or simply use one or several "unknown word tokens" – BitPar and LoPar use two unknown word distributions for upper- and lowercase words, and Petrov *et al.* (2006) reduce unknown words to a string based on the occurrence of upper-/lowercase letters, numbers, and other characters.

Dubey (2005) uses the suffix-based model introduced by Brants (2000) to yield a tag distribution for unknown words.

Experiments with the default behavior of BitPar showed that using case to choose one of two tag distributions for unknown words was clearly inadequate, since it led to unknown words being chosen to serve as verbs in an otherwise nonsensical analysis.

The approach for the parser was determined by two main considerations: On one hand, the architecture of BitPar allocates lexical entries statically, which meant that using a discrete set of lexical entries would be preferable. On the other hand, only looking at suffixes was clearly not enough, since past participles, and *zu*-infinitives of verbs with separable prefixes, cannot be distinguished from their infinitive counterparts by considering suffixes, since the past participle is built by pre- or infixation (consider *aufgeben/VVINF*, *aufgegeben/VVPP* and *aufzugeben/VVIZU*, which share the same suffix *-en* but would be easy to tell apart for anyone even vaguely familiar with German). Corresponding regularities would be rather easy to encode in regular expressions; However, an efficient mechanism for the combination of regular expressions is necessary to take advantage of them: On one hand, it is desirable to allow the system to combine regular expressions into more informative classes, on the other hand combining *all* regular expressions would yield an exponential number of cases and hence a massive sparse data problem. Manual combination of regular expressions into more informative criteria would be possible but labor-intensive and tedious. A good solution for this seemed to be to use a decision tree learner to identify highly informative feature combinations. I used dti, a package for decision tree induction written by Christian Borgelt[8], to grow a decision tree; the leaves of the decision tree would then be used as lexical entries for unknown words.

The current version uses 42 regular expressions that are used to induce a decision with a number of leaves ranging from 79 (for the original plain treebank grammar) to 111. The use of the decision tree for unknown words by itself yields a visible improvement (between 0.8% and 1.2% across the domains considered in Versley 2005) even in the case of the unmodified treebank grammar.

Using morphological analyses from SMOR While the approach based on regular expression heuristics eliminates some of the problem of unknown words being assigned implausible part-of-speech tags, a better handling of unknown words remains desirable since this problem becomes much more serious when applying the parser to different genres where the vocabulary does not overlap as much as it does within texts from the taz newspaper.

A significant amount of work going into hand-crafted parsers is in fact due to the extension of the parser's lexicon, and it is easy to see that using handcrafted knowledge on top of the information read off the treebank can potentially improve the parse quality if it is done in a robust way. However, the information we get from

[8] `http://www.borgelt.net/dtree.html`

a morphological analyser such as SMOR (Schmid *et al.*, 2004) is non-probabilistic and the analyses have a different form, which means that (i) it is necessary to remap the information from SMOR to our tagset, which is an extension of the Stuttgart-Tübingen tagset (STTS; cf. Schiller *et al.*, 1995) that includes some morphological information. Mapping SMOR's analyses to the extended-STTS tags involves some manual work but is not in itself new or exciting. It is also necessarty (ii) to introduce the predictions from SMOR into BitPar's parsing model in a suitable fashion.

The most common method, as used by Schulte im Walde (2000) and others, is to include the analyses for unknown words in the lexicon beforehand with a frequency of 0. This is not really satisfying since it does not allow to parse unlimited text, and it also seems to discard the probabilistic information that certain analyses for a word are more prominent than others.

The best approach to this was to use tag filtering: if an unknown word is found in SMOR, remove the tags that do not match any of those that were predicted by SMOR by deleting the corresponding bits in BitPar's chart. This allows to use the decision-tree-based prediction component to predict lexical items outside SMOR's lexicon (including rare and/or misspelt words), and allows to use both the probabilistic information from the decision-tree-based predictor and the categorical information from SMOR in the cases where an unknown word is covered by it.

Selective Markovization Markovization is a transformation that splits up rules and discards some of the context information.[9] It is usually seen as instrumental in reducing the sparse data problems that occur with long rule expansions. Schiehlen (2004a), however, contends that markovization, while improving constituent-based quality measures, yields a deterioration in terms of dependency-based quality measures.

For the TüBa-D/Z, the use of markovization is less obvious than for the NeGra/Tiger annotation scheme because of the TüBa-D/Z's more hierarchical structuring. However, the grammar with the improvements stated above still has a sparse data problem with **long coordinations** on one hand, where large numbers of conjuncts – the TüBa-D/Z, for example, contains conjunctions with up to ten conjuncts[10] – are strung together. Similarly, sentences that have many fragments -

[9] Normal rule binarization replaces parts of the right-hand side of an expansion by intermediate expansions that still carry the same information, e.g. replacing [NP a/DT long/JJ long/JJ story/NN] by [NP a/DT [JJ-JJ-NN long/JJ [JJ-NN long/JJ story/NN]]]. Markovization only keeps part of that information, resulting, for example, in an expansion such as [NP a/DT [JJ- long/JJ [JJ- long/JJ story/NN]]]. The grammar induced from the markovized expansion would also allow to generate [NP a/DT [JJ- long/JJ [JJ- long/JJ [JJ- long/JJ story/NN]]]] which neither the original grammar nor the normal rule binarization would allow.

[10] TüBa-D/Z, sentence 8491 (6 ADJX conjuncts):
Im festen Glauben, zu einer geheimen Gesellschaft zu gehören, die den unbestechlichen Blick auf den Mainstream gepachtet hat, produzieren sie Bücher, die keiner lesen will, [schwierige, experimentelle, unverständliche, verknotete und verfilzte] Texte in mausgrau klopapiernen Büchern, Bleiwüsten ohne Punkt, Komma und Absatz.
sentence 8996 (5 MF conjuncts):

usually enumerations rather than complete sentences – are also problematic and the parser would construct nonsensical analyses. To solve the problem of data sparseness in these cases, *selective* markovization seems to be the right solution: using markovization in those rule expansions where sparsity would otherwise be a big problem, but where the rules usually are very regular.

For example, the phrase

(6.13) [ADJX$_\varepsilon$ *entweder* [ADJX$_\varepsilon$ *lächerlich*] *oder* [ADJX$_\varepsilon$ *legendär*]]

(where ADJX$_\varepsilon$ simply means an adjective phrase with an uninflected adjective head) yields an expansion of length four:

$$\text{ADJX}_\varepsilon \rightarrow \text{KON}_1\ \text{ADJX}_\varepsilon\ \text{KON}\ \text{ADJX}_\varepsilon$$

In the transformed treebank, the phrase would be binarized to a subtree like the following:

(6.14) [ADJX$_\varepsilon$ *entweder*
[ADJX$_\varepsilon$/KON$_1$ [ADJX$_\varepsilon$ *lächerlich*]
[ADJX$_\varepsilon$/ADJX$_\varepsilon$ *oder* [ADJX$_\varepsilon$ *legendär*]]]]

which would yield multiple shorter expansions:

$$\begin{aligned}
\text{ADJX}_\varepsilon &\rightarrow \text{KON}_1\ \text{ADJX}_\varepsilon/\text{KON}_1 \\
\text{ADJX}_\varepsilon/\text{KON}_1 &\rightarrow \text{ADJX}_\varepsilon\ \text{ADJX}_\varepsilon/\text{ADJX}_\varepsilon \\
\text{ADJX}_\varepsilon/\text{ADJX}_\varepsilon &\rightarrow \text{KON}\ \text{ADJX}_\varepsilon
\end{aligned}$$

The shorter expansions occur more frequently, and result in reduced data sparsity, while the first-order Markov assumption still retains enough context that no frequent mis-parses occur.

For **middle fields**, where a great variety of adjuncts can come between the arguments and which therefore would be another candidate for markovization, first-order markovization would clearly result in a loss of useful context. A good intermediate solution seems to be to (i) break the middle field into parts which contain one argument each and (ii) mark these parts with the grammatical role of the preceding argument, so that only linguistically viable generalizations result.

As an example, consider the following example sentence (6.15)[11] where we have a middle field with a subject as well as dative and accusative objects.

Nach Ländern kommen [2.160 der neuen Soldaten aus Sachsen, 1.640 aus Thüringen, 1.610 aus Sachsen-Anhalt, 1.440 aus Brandenburg und 1.100 aus Berlin].
sentence 1946 (10 NX conjuncts):
Dann geht es in sieben Etappen a 160 Kilometer über [Olpe, Kassel, Weimar, Zeitz, Bad Lausick, Döbeln, Bad Schandau, Neustadt, Lübbenau und Storkow] nach Berlin.
sentence 11720 (10 NX conjuncts with 2 coordinations):
Preisträger sind die Berliner Filmtheater [Balazs, Bali, Central, Eiszeit, Filmbühne am Steinplatz, Filmkunst 66, fsk am Oranienplatz, Hackesche Höfe und Lichtblick-Kino sowie das Concerthaus Brandenburg].

[11] TüBa-D/Z, sentence 306.

(6.15) Wenn tatsächlich mal ein Junge mit ihr ausgehen will, läßt [$_{\text{MF}}$ [$_{\text{NX.ON}}$
If really once a boy with her go out wants, let
sie] [$_{\text{NX.OD}}$ sich] lieber [$_{\text{NX.OA}}$ das Geld für Drinks, Kinokarten und
she herself rather the money for drinks, cinema tickets and
Essen] vorher in bar] ausbezahlen.
food beforehand in cash disburse.
Whenever a boy actually wants to go out with her, she rather lets him disburse the money for drinks, cinema tickets, and food in cash.

The middle field of the matrix sentence yields an expansion with a right-hand side with six symbols:

$$\text{MF}_{\langle \text{ON,OD,OA} \rangle} \rightarrow \text{NCX}_n \ \text{NCX}_d \ \text{ADVX} \ \text{NX}_a \ \text{ADVX} \ \text{PX}$$

In the markovized version, the following single expansions would be generated:

$$\begin{aligned} \text{MF}_{\langle \text{ON,OD,OA} \rangle} &\rightarrow \text{NCX}_n \ \text{MF}_{\langle \text{ON} \| \text{OD,OA} \rangle} \\ \text{MF}_{\langle \text{ON} \| \text{OD,OA} \rangle} &\rightarrow \text{NCX}_d \ \text{MF}_{\langle \text{ON,OD} \| \text{OA} \rangle} \\ \text{MF}_{\langle \text{ON,OD} \| \text{OA} \rangle} &\rightarrow \text{ADVX} \ \text{NX}_a \ \text{ADVX} \ \text{PX} \end{aligned}$$

Since the number and kind of verb adjuncts can vary considerably, and the transformed representation additionally distinguishes between base noun phrases (NCX) and non-base noun phrases (NX), this strategy helps avoid data sparsity issues, in addition to the topological field annotation, which also provides a useful generalization here. In comparison, a similar strategy for the NeGra or Tiger treebanks, which do not have a MF node, would either be very difficult or, for approaches like Dubey's (2005), would be prone to producing linguistically unsatisfying analyses with multiple subjects or multiple finite verbs.

An Example Figure 6.3 shows an example of the tree transformations used for deriving the PCFG grammar: verbs are subcategorized by auxiliary (VAFIN:haben) or by argument valence (VVPP:a), whereas articles and noun contain morphological information. Noun phrases have case information (NCX_{n} and NCX_{a}), and the complete sentence is marked as a V2 sentence (SIMPX_{v2}). The middle field is marked as containing a subject (ON) and an accusative object (OA), and contains an intermediate node ($\text{MF}_{\langle ON \| OA \rangle}$) due to the markovization of the middle field.

6.2.2 Extracting Grammatical Relations

Models such as the ones by Hindle (1990) or Lin (1998a) use grammatical relations, whereas the parser generates constituent structures. To extract these grammatical relations, several steps are necessary: One is to reconstruct the *edge labels* of the parse tree, which contain information about the relation of a constituent node to its parent node (for example, ON for subjects and OA for accusative objects). The

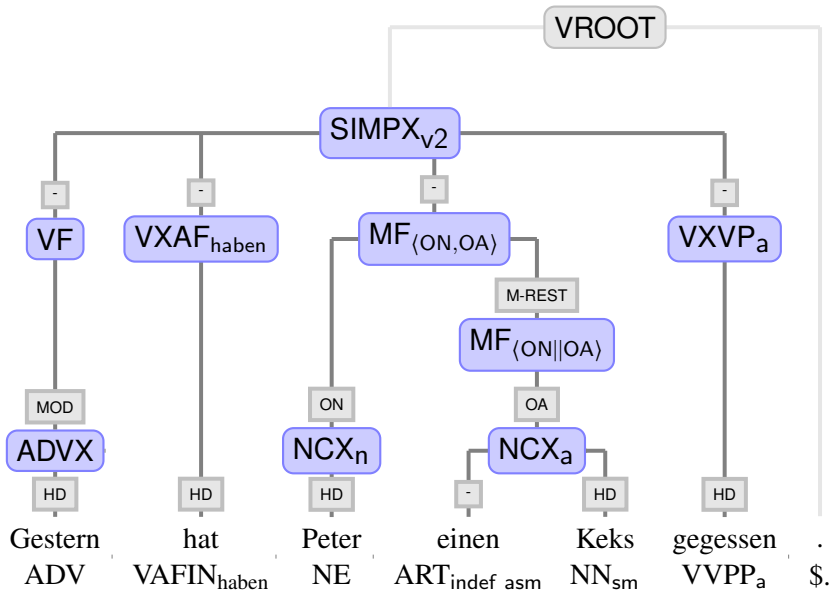


Figure 6.3: Annotated Treebank Grammar representation (*final*)

tree of constituents-with-edge-labels is then converted to a dependency representation. Finally, the grammatical relations are read off from the dependencies, with additional processing to allow for passives and copula sentences.

An alternative to this approach consisting of a dependency representation in a first step and extraction of grammatical relations in the second would be a more direct approach that aims at extracting grammatical relations directly from the constituent representation. Such an approach has been proposed by Forst (2003), who converts trees of the TIGER treebank into f-structures. Forst's unification-based approach creates a richer representation than a conventional single-parent dependency annotation and can cover phenomena such as subject and object control, but is less flexible with regard to the relation between constituents and grammatical relations. As single-parent dependencies are conceptually simpler and an annotation scheme for single-parent dependencies readily existed (Foth, 2003), the easier evaluation of dependencies made the two-step approach more appealing.

After the edge labels are reconstructed by selecting the most frequent edge labeling corresponding to a given PCFG rule, the constituent representation is transformed to dependencies by a combination of head projection (Magerman, 1995), recovery of verb complements (Kübler and Telljohann, 2002), and post-correction of dependencies to reduce the divergence between the conversion result and the CDG target annotation, similar to Daum *et al.* (2004).

As we aim to stay close to the CDG annotation scheme, which attaches objects to the full verb, and the full verb is not necessarily the finite verb of the sentence,[12]

[12] Attaching subjects to the finite verb, and objects to the full verb, is preferable for dependency

the dependency extraction process on clause level keeps track of several lists of dependents:

- `lk_deps` contains a list of the arguments to the finite verb, which includes subjects, complementizers, and prepositional modifiers in the pre-field, which the CDG grammar attaches to the finite verb.
- `vc_deps` contains all non-subject verb arguments
- `fin_verb` contains the finite verb (or verbs, in case of annotation errors)
- `aux_deps` contains all non-finite verbs, as well as verb particles.
- `konj_deps` contains a list of additional conjuncts in the case of field coordination. In field coordination, the first FKONJ node is counted as part of its containing clause, whereas the subsequent children are connected as conjuncts.

After collecting the children of a clause, the elements from `lk_deps` and `vc_deps` are chained up via AUX edges (or, respektively, an AVZ edge for verb particles), the `fin_verb` dependants are linked to the finite verb and the `vc_deps` dependants are linked to the full verb (i.e., the verb that comes last in `aux_deps`.

A few labels cannot be accurately deduced by the label rules:

- The first is the distinction between genitive postmodifiers (`GMOD` in the CDG grammar) and other nominal postmodifiers (where the CDG grammar uses the `APP` label). The TüBa-D/Z only has an unspecific head/non-head label (‘`-`’) for both cases, and a number of heuristics including the case of the modifier and the presence of a determiner is used to assign a relation.
- The second one is TüBa’s `OS` label: it is used both for complementizer subclauses and for verb-second embedded clauses. The postprocessing routine checks for a possible complementizer and assigns an according grammatical function.

As mentioned above, nonprojective dependency structures, as they occur in postposed relative clauses, are encoded either in the relative clause’s edge label or in a secondary edge. Secondary edges are not recovered by the PCFG parser, and the edge label recovery process does not have enough information to recover the edge labels here (since the rule in all cases would be something like `NF` → `R-SIMPX`, the context is not sufficient), but it should in principle be possible to construct a separate component for reattaching these.

A discussion of the conversion quality of this approach can be found in Versley (2005). In a comparison to a hand-annotated sample of 300 sentences, verb arguments such as subjects and objects are converted in near-perfect quality (to 99%

parsing, due to the fact that subject agreement has to be checked on the finite verb, and argument requirements have to be checked against the full verb.

	SUBJ	OBJA	OBJD	PP	GMOD	LAS	UAS	sec./sent[a]
WCDG[b]	0.90	0.80	0.65	0.78	0.86	0.88	0.90	76.0
S/R	0.72	0.46	0.02	0.64	0.65	0.74	0.80	0.003
vanilla	0.69	0.31	0.04	0.66	0.39	0.74	0.82	0.75
final	0.88	0.82	0.60	0.71	0.74	0.86	0.88	1.00

[a] Speed measured on a 2.66GHz Intel Core 2.

[b] Accuracy figures for WCDG were determined against manually corrected dependency parses from a 300-sentence sample from TüBa-D/Z (Versley, 2005), all others against the development set of the ACL 2008 Parsing German shared task.

Table 6.1: parsing accuracy on TüBa-D/Z

and 95%, respectively), whereas adverbs and prepositional phrases are treated differently by the annotation guidelines of TüBa-D/Z and the CDG scheme. Altogether, the conversion has an accuracy of about 93%.

The conversion to grammatical relations directly uses the dependency edges in most cases. The handling of subjects, however is a bit more complex, because the subject is not always attached to the full verb, but also because of passives and predicative constructions.

- For predicative constructions involving the verbs *sein* (be) or *werden* (become), a `PRED` relation is extracted between the subject and the predicate.
- Eventive passives (a) and *zu*-passives (b) are identified by the combination of auxiliary and the form of the dominated verb; past participles with *sein* auxiliaries are checked against the CDG lexicon to discriminate between verbs normally having a *sein* auxiliary (d) and stative passives (c).

(6.16) a. Der Rasen wird gemäht.
The grass is being mowed.

b. Der Rasen ist zu mähen.
The grass is to be mowed.

c. Der Rasen ist gemäht.
The grass is mowed / has been mowed.

d. Der Rasen ist gewachsen.
The grass has grown.

Parsing Experiments Table 6.1 compares the parsing accuracy on different kinds of dependency edges, labeled and unlabeled attachment accuracy, and speed for various parsers discussed here. For the WCDG parser, the comparison was made against a hand-corrected sample of sentences to ensure a fair comparison despite differences in the annotation scheme. Neither Foth's shift-reduce parser (*S/R*) nor the unmodified treebank grammar (*vanilla*) achieve adequate results in

the recognition of accusative and dative objects; furthermore, PCFG parsing with the unmodified treebank grammar is inaccurate when it comes to postmodifying genitives. The WCDG parser (tested using unsupervised PP attachment but not incorporating predictions by the shift-reduce parser) reaches good accuracy on subjects and accusative objects and performs adequately on the more difficult dative objects, but is relatively slow.

The parsing approach presented here (designated as *final* in the table) reaches an accuracy on verb arguments that comes close to the WCDG parser's, but is considerably faster. Using the combination of WCDG parser and prediction by the shift-reduce parser, as presented in Foth and Menzel (2006), would reduce the speed difference to a factor of three to five, which is still considerable.

As seen in the parse of figure 6.4, the PCFG trained on the transformed treebank achieves accurate identification of both subject and accusative objects. Attachment of prepositional phrases is still a problem, which however is difficult to solve using unlexicalized parsing.

Some of the approaches presented in section 5.3.3 would conceivably produce results that are faster or more accurate, but were not available in 2005/early 2006, when the experiments described in this section were done. The Berkeley Parser (Petrov *et al.*, 2006) is now (as of 2010) widely recognized to be a good choice for parsing of languages other than English, performing well on German (Petrov and Klein, 2008b), Italian (Lavelli and Corazza, 2009) and French (Seddah *et al.*, 2009), however its suitability for parsing languages other than English was only asserted in (Petrov and Klein, 2007).

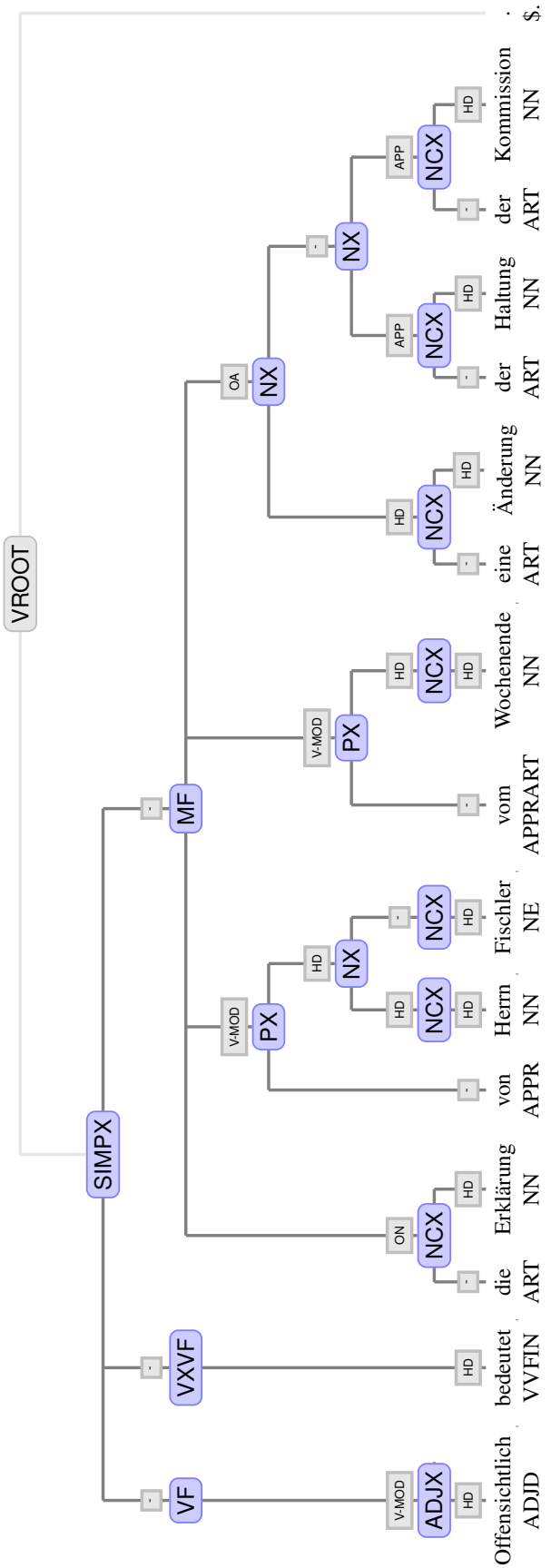


Figure 6.4: Parse with better grammatical relations (*final* grammar)

Of the current dependency parsers, both the dynamic-programming based MSTParser (McDonald, 2006) and Nivre's SVM-based version of MALTParser (Hall *et al.*, 2006) reached good results in the CoNLL 2006 evaluation (Buchholz and Marsi, 2006). Other transition-based parsers such as the one of Attardi and Ciaramita (2007) did significantly worse, which suggests that extensive feature tuning is necessary to achieve useful performance using transition-based parsers.

6.2.3 Compound splitting

In English, *Minister of Interior* has the head word *minister*, which is a more general concept than *minister of interior*. Thus, analytic compounds give us one step of generalization for free in a task where data sparsity may create problems otherwise. In German, *Innenminister* (minister of interior), *Aussenminister* (minister of foreign affairs) or *Umweltminister* (minister of environment) all have single, different word forms, which contributes to a much larger type/token ratio in German and may create a severe sparse data problem. Using compound splitting, we can use the head part of a synthetic compound (*Minister* in this case for all the compounds) to eliminate these data sparsity problems by merging several word forms that are (most probably) related.

Compound splitting has long been recognized as a very useful component even for natural language processing tasks that traditionally prefer shallow processing mechanisms, for example in machine translation (Niessen and Ney, 2000), and information retrieval (Monz and de Rijke, 2001).

The simplest conceivable approach to splitting compounds, used in the WCDG parser to improve the coverage of the lexicon (Foth, p.c.), is to simply select the longest suffix that corresponds to a known word. This is error-prone in some cases, where words correspond to frequent non-word suffixes (for example in cases like *Abstraktion⇒Ion*). A more accurate approach than simply taking any valid suffix would be to use a morphological analyzer, as done in the PARGRAM LFG grammar (Butt *et al.*, 1999), which reduces the impact of misanalyses due to non-word suffixes, but yields multiple analyses which have to be disambiguated in some way, in addition to being unable to cover words that have not been modeled in the morphological analyzer.

Larson *et al.* (2000) use a purely data-driven approach that tries to find possible word junctures by considering how many word tokens in the training corpus have a prefix/suffix that is common with this letter. They then use the differences of these counts to get an indication of left/right morpheme boundaries, and insert left/right boundary candidates at local maxima. Splits are then inserted at positions identified both as a left- and right-boundary candidate, with an additional filtering step that rejects any split that results in non-pronounceable parts.

The approach of Monz and de Rijke (2001) is completely deterministic and is based on a lexicon: their splitting algorithm recursively tries to decompose the word into an item from the lexicon that is as small as possible, possibly a glueing-*s*, and a part that can either be decomposed further or is an item from the lexicon.

Brown (2002) uses a parallel corpus to find cognate pairs of synthetic (on the German side) and analytic (on the English side) compounds. Assuming a 1:2 split, he first constructs the best matching prefix and suffix of the compound using an alignment of identical or approximate matching letters, and subsequently adjusts the boundaries so that an exact division between the two words is found. In a subsequent steps, compounds based on one cognate and one translated word are considered.

Koehn and Knight (2003) use a two-phase approach to learn compounds from a parallel corpus: In the first phase, they consider all possible decompositions of a noun into several known words, with possible intervening linking morphemes. These are then ranked according to the geometric mean of their word frequencies.

The second phase uses a parallel corpus to estimate both the likelihood that the German compound really corresponds to two English words and translations for prefixes of compounds, which may differ from the original words (consider *Grundrechte*, which would be translated as *fundamental/basic rights* rather than *foundation/reason rights* which could be suggested by the translation of the noun *Grund*).

Schiller (2005) uses a finite-state morphological analyzer in addition to multiplicative weights that provide an ordering of the analyses. Starting from a uniform segment weight of 0.5 (which corresponds to giving the best score to the alternative(s) that have a minimal number of segments), she increases the weights of segments that have a high chance of being part of the right analysis for their surface form:

$$\text{weight}(lex : srf) = 0.5 + 0.2 * \frac{\text{freq}(lex : srf)}{\text{freq}(: srf) + 1}$$

The factor of 0.2 is chosen so that even with varying segment weights, a solution with fewer segments would always be preferred over a solution with more segments. Note that Schiller's approach, in addition to being limited to the coverage of the morphological analyser, needs annotated training data, while the other approaches above are essentially unsupervised.

Marek (2006) uses a weighted finite state approach such as Schiller's, but without a full-blown morphological analyzer. He uses a lexicon and a shallow predictor for linking morphemes to construct possible analyses, which are also disambiguated by hand to create the gold-standard data used for later training and evaluation.

Marek then uses weights that approximately correspond to a generative model that first generates a word and then a linking morpheme: Words get a weight that corresponds to their probability of occuring as part of a compound:

$$W(w) = -\log\left(\frac{C(w)}{N}\right)$$

Linking morphemes are associated with a weight that corresponds to their probability of occurrence, given a particular word:

$$W(l) = -\log\left(\frac{C(l_w)}{C(w)}\right)$$

The model is completed by using words with similar suffixes to predict the linking morpheme distribution for words that do not occur in non-final position, and a wildcard pattern for unknown words (using prefixes and suffixes of known words, together with a constraint that the unknown segment has to contain both consonants and vowels). Unknown words are then weighted using the probabilities of the respective prefixes.

Of the approaches mentioned before, only Monz and de Rijke (2001) and Koehn and Knight (2003) provide both a gold-standard-based and an application-oriented evaluation, whereas Larson *et al.* (2000) provide only an application-oriented evaluation and Schiller (2005) and Marek (2006) provide only a gold-standard evaluation.

While Monz and de Rijke arguably have the worst compound splitting algorithm, achieving 61% recall and 50.6% precision (against close to 99% in the case of Koehn and Knight, Schiller or Marek), they still report an improvement on the information retrieval task. This may be related to the way they integrate the compound information in the IR system, essentially *adding* the compound parts to the original documents and leaving the unsplit compounds in place.

Koehn and Knight report increases in translation quality for their phrase-based statistical machine translation system, for all the compound splitting methods, but note that the simpler monolingual corpus-based method gives a more noticeable increase than the more elaborate methods. A possible explanation for this would be that the phrase-based SMT system is able to learn and correct the oversplit compounds.

Criteria for Compound splitting How should a compound splitter for distributional similarity look like? First, since we aim for similarity and not for mere association, we are mainly interested in the heads of compositional compounds: while an *Umweltminister* (Minister of Environment) is a kind of *Minister* (minister), an *Augapfel* (eyeball) is not a kind of *Apfel* (apple), and some compounds may be somewhere in-between, as a *Gasthaus* (pub) is a kind of *Haus* (house), but different enough that it should not be treated as one (*Gasthaus* allows a polysemous reading of pub-as-business, whereas *Haus* does not have an organization reading).

While noticing the compositionality of nouns would require the distributional similarity metric that we need them for in the first place, it can be said in a good approximation that rare compounds, where we need compound splitting most from a sparse data point of view, are mostly regular and compositional, whereas noncompositional compounds tend to be (relatively) more frequent. Thus, the application in mind (distributional similarity measures) would be served best with a frequency-sensitive algorithm for compound splitting.

634	Haupt-	407	Partei-	371	Bundes-
522	Frauen-	404	Stadt-	370	Polizei-
500	Gesamt-	403	Welt-	363	Finanz-
485	Kunst-	403	Groß-	363	Bau-
474	Kultur-	400	Riesen-	356	Wahl-
470	Film-	398	Theater-	356	Regierung-
462	Kinder-	395	Fernseh-	355	West-
430	Wirtschaft-	387	Arbeit-	354	Verkehr-
430	Medien-	381	Ost-	351	Staat-
420	Umwelt-	372	Wasser-	349	Sonder-

Table 6.2: Most frequent prefixes

Since we are only using the last part of the word, requiring the split to use only covered lexical items would also make less sense than using a best-effort heuristic matching to find a plausible head.

The compound splitting algorithm that I use first collects frequent compound *prefixes* to then collect evidence for a set of frequent compound heads (the *base nouns*). In absence of other evidence, these base nouns (and, to a lesser extent, prefixes from the list) are then preferred for the shortening of infrequent compounds.

The first script collects all the possible splits of rare words ($n < 5$) into a prefix and a non-rare ($n >= 20$) word. The possible splits are then aggregated by type (to identify highly productive prefixes). A prefix is then considered valid if it fulfills the following criterion:

$$\frac{\mathrm{C}(\mathit{prefix}) \cdot e^{\mathrm{length}(\mathit{prefix})}}{\mathrm{C}_{\mathrm{total}}} > 5$$

where $\mathrm{C}(\mathit{prefix})$ is the type count of the prefix, the $e^{\mathrm{length}(\mathit{prefix})}$ term prefers longer prefixes (very short prefixes trivially occur with a large number of words), and $\mathrm{C}_{\mathrm{total}}$ is the total type count of rare words.

Collecting these frequent prefixes is especially useful since many of the prefixes (including *Haupt-*, *Sonder-* or *Finanz-*) do not correspond to frequent nouns.

Another script creates a statistic of suffixes (potential compound heads) according to whether they co-occur with previously identified *good* prefixes. A suffix is identified as a potential compound head if it either always occurs with a good prefix or if it occurs with 'bad' prefixes less than 7 times, and in less than 10% of cases. The compound heads thus identified may still contain compounds such as *Ostgebiet*, *Ostpreußen*, or *Ozonloch*, but since I wanted to reduce the impact of low-frequency items rather than provide a complete morphological analysis, this is not necessarily bad.

A final script is used to provide the actual mapping from raw lemma forms to the forms used for the similarity calculation. `compound3.pl` performs the following steps:

- the *-in/-innen* suffix of feminine forms involving certain suffixes is removed (ex.: *Friseurin* ⇒ *Friseur*).
- Any word with a count of more than 100 is not modified further.
- Any name that occurs in a gazetteer (of person and location names) and is not a base noun is normalized to the category assigned by the gazetteer.
- If the word has a base noun as a suffix, return the longest such suffix.
- Find the longest suffix such that the suffix is a word that has been seen more than 20 times and the prefix is a *good* prefix. If such a suffix exists, return it.
- If the word has occurred more than 20 times in the corpus, simply leave it unmodified.
- Otherwise, look for the longest suffix that occurs more than 20 times in the corpus.
- Normalize rare abbreviations (with three capitalized letters) to the string "*(ABK)*".

6.2.4 An Efficient Implementation of Distributional Similarity

Schütze (1992), explaining that he used singular value decomposition for his word sense disambiguation experiments, says that "*If 20,000 words were to be represented with 5000-component vectors each, a 100-megaword memory would be required, and any application program, for instance for word sense disambiguation, would be prohibitively slow*". While now, fifteen years on from Schütze's use of supercomputing for distributional similarity, using hundreds of megabytes of memory does not scare the seasoned computational linguist, it has to be said that scalability is still a significant issue for the integration of distributional semantics measures in a coreference system.

In particular, while it is possible to limit the amount of data that has to be processed and loaded in memory to a much smaller amount in the case of small clustering evaluations, or in the case of the TOEFL dataset used by Landauer and Dumais (1997) and many others, use in a coreference system requires that a large, open set of lexical items be stored.

A first implementation using pure Python data structures – representing the set of co-occurring items for words as a database-backed hash of hashes – was unusably slow and subsequently abandoned for an approach that stores sparse vectors in compact form as dense number arrays using the NumPy package.

For the efficient creation of these count vectors, extracted relations are first stored in a hash-based data structure, and then converted into a dense form when the hash has grown to a certain size, progressively merging these partial sparse matrices to create compact structures of growing size until all of the data has been

processed. This technique of creating static data structures and always merging two similar-sized structures into one that is twice as large ensures that only $O(n \log n)$ time is consumed while still offering the space savings of a compact, static data representation. This static representation can then be loaded as a memory-mapped file, with the advantage that operating system facilities take care of loading (and, in cases of memory scarcity, unloading) the count matrix into memory and the possibility for multiple processes on the same machine to share the memory pages containing the count matrix.

The C++-based computation of similarity metrics (accessing the NumPy data structures through the Python/C interface) is then fast enough to be used on the fly, with the exception of most-similar item lists, where a small number of most-similar items (250) is cached for each word to avoid the cost of recomputing the similarity measures between that word and all 20,000 other entries each time.

6.2.5 Similarity Measures Induced from the taz Corpus

Using all sentences with a length of between 5 and 30 tokens from the years 1986 to 1998 of the taz corpus (which gives a total of over 8 million sentences, of which 32 thousand, or 0.3%, could not be parsed), I extracted count matrices for subject and object relations, modifying adjectives, as well as both modifying and modified items in genitive postmodification and postmodifying prepositional phrases with *in*, *von*, and *bei*. The resulting similarity measure is named **Lin98** in the following discussion.

In order to assess the design choices for such a similarity measure, I also include two other measures which are based on unsupervised learning, but crucially differ from Lin's distributional similarity measure in one or more of the design decisions: One is the relation-free distributional model of Padó and Lapata (2003), which defines neighbourhood over dependency graphs (cf. the earlier discussion on p.153), which uses mutual information weighting and a Jaccard-like vector similarity function like Lin's measure, but does not distinguish between different dependency relations. The second is a version of Garera and Yarowsky's (2006) model which is based on first-order association between definite descriptions and plausible same-head antecedents (cf. the earlier discussion on page 6.1.3).

The three approaches make very different use of syntactic preprocessing: while the Lin approach presupposes accurate identification of grammatical relations, partially labeled parses are sufficient for Padó and Lapata's approach, and the Garera and Yarowsky approach merely requires sentences and chunks. Because not all sentences were processed with the PCFG parser, unlabeled dependency trees for the whole taz corpus were created using Foth's shift-reduce parser. Following Padó and Lapata (2003), the syntactic neighbourhood relation was defined over all within-clause dependency relations (modifying adjectives, adverbs, genitive modifiers, and verb adjuncts and arguments). The resulting similarity measure is referred to as **PL03** in the following discussion.

For the reimplementation of Garera and Yarowsky's association measure, the chunk annotation from TüPP-D/Z (Müller, 2004b) was used. From each sentence in a document, all noun chunks were extracted along with their heads. Following Garera and Yarowsky (2006), for any definite noun phrase, a noun chunk with the same head was searched for in the previous sentences. In the case where no same-head antecedent was present, antecedent candidates from the previous two sentences were collected and their heads paired with the head of the definite noun phrase. Using mutual information weighting results in an association measure that directly reflects potential anaphor-antecedent relations in the text collection. For this association measure, I will use the shorthand name **TheY**.

As a slight variant of the Garera and Yarowsky measure, I will also consider a variant that uses the G^2 (likelihood ratio) measure of Dunning (1993), which I will call **TheY:*G*^{2}**. This second measure has a stronger preference towards high-frequency items: While the mutual information statistic measures the size of the difference between actual and predicted counts, the G^2 measure is related to statistical significance and grows with the sample size.

The main purpose of these similarity measures is the comprehensive evaluation in the antecedent selection task of section 6.4. However, a more impressionistic evaluation can provide a kind of reality check here and help to assess the type of antecedents that can be found by these unsupervised measures. Concrete objects – vehicles, animals, food – can usually be classified relatively well (cf. the results of Almuhareb and Poesio, 2005b), whereas abstract objects are more difficult. Therefore, table 6.2.5 concentrates on three high- to mid-frequent abstract nouns that commonly occur in definite descriptions: *Krieg* (war; rank 349 according to lemma frequency in TüPP-D/Z), *Land* (country; rank 119) and *Medikament* (medical drug; rank 3203).

Looking at the table contents, we see that both Lin's relation-based similarity measure and Pado and Lapata's relation-free measure yield many terms that are substitutable with the term itself; however, Pado and Lapata's measure also retrieves *Behandlung* (treatment) which is semantically distant. Many of the terms found are not exact synonyms, but either substitutable in some contexts (Land/*country* vs. Bundesrepublik/*federal republic*) or semantically related but not very close, such as *Medikament* (medical drug) and *Lebensmittel* (foodstuff), which can both be ingested, but have different functions and properties otherwise.

Garera and Yarowsky's TheY measure is asymetric, which results in the retrieval of *RU* (for RU 486) and *Viagra* as instances of *Medikament*. On the other hand – and this is plausible since similar techniques have also been proposed for associative bridging – this metric retrieves a large number of related but non-substituting terms such as *Pharmakonzern* (pharmaceutical company) for *Medikament*. Using the mutual information statistic for association also results in many rare words being ranked very highly, such as *Kemalismus* (Kemalism) for *Land* (country) – presumably, all mentions of Kemalism occur in a context where its development in a country is discussed.[13]

[13] *Kemalism* is the name for the laicistic system of beliefs introduced by Kemal Atatürk in post-Osman Empire Turkey.

As can be seen in the column labeled *TheY:*G^2, substituting mutual information with the G^2 (likelihood ratio) measure of Dunning (1993) – an association statistic that puts a stronger emphasis on frequent associates – we see that the modified measure retrieves more believable collocates, which however are not substitutable – pairs such as *Krieg-Stadt* (war-city), *Land-Regierung* (country-government) or *Medikament-Patient* (drug-patient) dominate the list of most close associates.

Geffet and Dagan (2004) propose an approach to improve the quality of the feature vectors used in distributional similarity measures: instead of weighting features using the mutual information value between the word and the feature, they propose to use a measure they call *Relative Feature Focus*: the sum of the similarities to the (globally) most similar words that share this feature.

By replacing mutual information values with RFF values in Lin's association measure, Geffet and Dagan were able to significantly improve the proportion of substitutable words in the list of the most similar words. As can be seen in the table, RFF, like Lin's measure, has a higher proportion of semantically similar, rather than simply associated, terms. However, the quality of the extracted most-similar words is not improved with respect to the original lists from the Lin measure. Subsequent experiments revealed that Geffet and Dagan's method introduces a greater focus on high-frequency words, which makes sense in the context of ontology learning, but is not as useful for the applicability in tasks like coreference resolution.

Finally, we can compare the results from these association metrics to the German LSA word space[14] which has been created and made publically available by Tonio Wandmacher. In the LSA-based word space, the nearest neighbours of *Krieg*(war) are *Balkan*, *Frieden* (peace), *Weltkrieg* (world war) and *Zivilbevölkerung* (civilist population), those for *Land* are *Bürgerkrieg* (civil war), *Bund* (federate government), *Brandenburg*, *Landeskasse* (state bank), *Hektar* (hectare), and those for *Medikament* (medical drug) are *Behandlung* (treatment), *Mediziner* (medical practitioner), *Nebenwirkungen* (side effects), *Patient* (patient) and *Virus*, indicating that we get thematic associations but not similarity.

[14] `http://www.cogsci.uni-osnabrueck.de/~korpora/ws/cgi-bin/HIT/LSA_NN.perl`

Lin98	RFF	TheY	TheY:G^2	PL03
		Krieg *(war)*		
Bürgerkrieg	Kampf	Kriegsführung	Golfkrieg	Krieg
civil war	*combat*	*warfare*	*gulf war*	*war*
Kampf	Bürgerkrieg	Kriegsfilm	Weltkrieg	Bürgerkrieg
combat	*civil war*	*war movie*	*world war*	*civil war*
Krise	Konflikt	Ostfront	Stadt	Konflikt
crisis	*conflict*	*east front*	*city*	*conflict*
Auseinandersetzung	Auseinandersetzung	Stellungskrieg	Serbe	Weltkrieg
dispute	*dispute*	*static warfare*	*Serb*	*world war*
Streit	Streik	Kriegsberichterstattung	Soldat	Schlacht
quarrel	*strike*	*media coverage of a war*	*soldier*	*battle*
		Land *(country/state/land)*		
Staat	Staat	Kemalismus	Regierung	Kontinent
state	*state*	*Kemalism*	*government*	*continent*
Stadt	Stadt	Bauernfamilie	Präsident	Region
city	*city*	*agricultural family*	*president*	*region*
Region	Landesregierung	Bankgesellschaft	Dollar	Stadt
region	*country government*	*banking corporation*	*dollar*	*city*
Bundesrepublik	Bundesregierung	Baht	Albanien	Staat
federal republic	*federal government*	*Baht*	*Albania*	*state*
Republik	Gewerkschaft	Gasag	Hauptstadt	Bundesland
republic	*trade union*	*(a gas company)*	*capital*	*state*
		Medikament *(medical drug)*		
Arzneimittel	Pille	RU	Patient	Arzneimittel
pharmaceutical	*pill*	*(a drug*)*	*patient*	*pharmaceutical*
Präparat	Droge	Abtreibungspille	Arzt	Lebensmittel
preparation	*drug (non-medical)*	*abortion pill*	*doctor*	*foodstuff*
Pille	Präparat	Viagra	Pille	Präparat
pill	*preparation*	*Viagra*	*pill*	*preparation*
Hormon	Pestizid	Pharmakonzern	Behandlung	Behandlung
hormone	*pesticide*	*pharmaceutical company*	*treatment*	*treatment*
Lebensmittel	Lebensmittel	Präparat	Abtreibungspille	Arznei
foodstuff	*foodstuff*	*preparation*	*abortion pill*	*drug*

highest ranked words, with very rare words removed

*: RU 486, an abortifacient drug

Lin98: Lin's distributional similarity measure (Lin, 1998a)

RFF: Geffet and Dagan's *Relative Feature Focus* measure (Geffet and Dagan, 2004)

TheY: association measure introduced by Garera and Yarowsky (2006)

TheY:G^2: similar method using a log-likelihood-based statistic (see Dunning 1993)
this statistic has a preference for higher-frequency terms

PL03: semantic space association measure proposed by Padó and Lapata (2003)

Table 6.3: Similarity and association measures: most similar items

6.3 Additional Semantic Features

This section presents two information sources for coreference resolution that occupy a middle ground between more precise approaches such as lookup of relations in GermaNet, and the more recall-oriented, less-precise methods based on unsupervised learning from text.

The first is a pattern-based approach similar to the one presented by Markert and Nissim (2005), which uses high-precision surface patterns in conjunction with the World Wide Web as a data source, which makes it possible to achieve useful recall at the same time.

The second is an approach that harnesses gazetteers and information from GermaNet to determine a coarse semantic class for a mention. The semantic class represents relatively coarse-grained information, but is relatively reliable and can be used to exclude implausible antecedents even when no other information is available.

6.3.1 Pattern search for German

To implement the pattern-based search for German, I do not use compound splitting, as compounds not only carry valuable information, but also may be more frequent than their compound heads. While the general algorithm of Markert and Nissim's (2005) pattern search idea is quite simple – take the heads of the noun phrases, use morphological generation to provide slot fillers, and use a search engine's web service to retrieve the count – there are still stumbling stones that make a high-quality implementation non-straightforward.

Let us consider a simple example:[15]

(6.17) a. Kilos und Fitneß sollen stimmen, denn Wesemann will [die Tour de France] gewinnen. (...)
Kilos and fitness have to be right, as Wesemann wants to win [the Tour de France].

b. Viele führen nun erneut den Vergleich mit [dem spektakulärsten Radrennen der Welt] im Mund.
Once again, many are using the comparison to [the most spectacular cycle race of the world].

We would like to confirm that *Radrennen* (cycle race) could indeed be a subsequent mention of *Tour de France* by generating a query like "*die Tour de France und andere Radrennen*". Note that (i) since *Tour de France* is a proper name, we are not merely using the head word *Tour*, but the complete name and (ii) since the original noun phrase contains a determiner, we include it in the search string even though it is not a common noun.

As a first step, we need to normalize the surface form to minimize the influence of case inflection and/or the strong/weak inflection in the case of deadjectival

[15] TüBa-D/Z, sentences 1494,1500

	X u. a. Ys	Ys wie X	Ys einschl. X	Ys insb. X	Ys außer X
Monsanto – Firma	30	243	—	—	—
Bremen – Stadt	22	231	5	2	3
Korsika – Insel	1	51	1	—	—
Magath – Trainer	1	31	—	—	—

	das Y X	die Ys X	X ein Y	X ist ein Y	X ist das Y
Monsanto – Firma	571	76	6	1	2
Bremen – Stadt	39.800	846	231	465	67
Korsika – Insel	674	130	66	142	17
Magath – Trainer	112	1	11	9	5

Table 6.4: Counts for selected patterns

nouns. For this purpose, we use SMOR (Schmid *et al.*, 2004), a finite-state morphology model for German, to normalize the nouns to nominative form and generate singular and plural versions. Using SMOR for generation results in spurious ambiguities, for example, generating the nominative singular form of *Radrennen* - bicycle race - results in 3 additional surface forms, *Räderrennen*, *Radsrennen*, *Radesrennen*. We then choose a form by preferring surface forms that are closer to the original string in terms of edit distance, more frequent globally (to allow generating the dissimilar "Länder" from the genitive "Landes", rather than the similar, and equally valid, but marked, "Lande"), and more similar frequencies (to prevent spurious matches by confusion with high-frequency words).

Using the pattern "X *und andere* Y_{pl}" (X and other Ys), the system can correctly resolve a few more instances (improving recall by a comparatively meager 1.2%), but the precision is much lower than with GermaNet, making it worse both in terms of recall and precision than the wordnet-based approach. Another pattern that Markert and Nissim mention, but did not include in their study, "Y_{pl} *wie* X" (Ys such as X), yields a higher gain in recall, as it is found more frequently, with a recall gain more than three time as high, but also better precision (54% for the part that is handled by the web-based resolver). However, even with this pattern, the achievable recall is still inferior to hyponymy-based lookup in a wordnet. The main reason for this seems to be that data sparsity is still a problem, more so for "X und andere Ys" than for the higher-recall pattern "Ys wie X".

The reason why low recall for a pattern presents not only a recall problem, but also a precision problem can seen in cases where one potential (but wrong) antecedent is found by the pattern, but the correct one is not, for example in *die Region* (the region) wrongly being resolved to *China* instead of *das Kosovo* because China occurs so much more frequently than the Kosovo that the pattern is seen with the former, but not with the latter.

Many cases where a wrong antecedent is found are due to this kind of error, for example the correct antecedent *Pioneer* being less known as a company dealing

in seeds (*Saatguthersteller*) than *Monsanto*, which is further away and the incorrect antecedent, *Bonn* being chosen instead of the lesser-known *Gummersbach* as antecedent for *die Stadt*, or *Hamburg* instead of the small city *Elmshorn* for *der Ort*.

Combining both patterns (i.e. choosing a potential antecedent if either pattern can be found) does not lead to increased recall, and precision is between the values for each pattern alone. Also, enlarging the sentence window (from 4 to 8 sentences) only yields more false positives, which suggests that the web counts capture different information than the hyponym search in GermaNet.

Harnessing the sheer size of the Web with low-recall methods still seems more difficult for German than for English, since German data is sparser by approximately one order of magnitude: Kilgrariff and Grefenstette (2003) give a factor of about 1:10 for the size ratio of German to English web indexed in AltaVista. For counts of the patterns that interest us, the numbers can vary wildly, between 4:1 for *Bremen*, and 1:250 for *Houston*. *"Monsanto and other companies"* yields 523 hits on google.com, while its German counterpart only yields 6; the higher recall pattern *"companies such as Monsanto"* yields 25,500 hits for English, while its German counterpart only gets 236 hits. The problem may be even worse for other languages like Dutch or Italian, which have an even smaller number of (native or non-native) speakers and consequently also a smaller amount of text that is available via web searches.

Given the computational cost of morphological regeneration and also the strict limit on search queries (the Google API has a limit of 1000 queries per day, while the Yahoo API has a larger limit of 5000 queries per day, but sometimes returns wildly inaccurate results for phrase queries), it is quite clear that only using web queries with shallow patterns is not a solution in itself. But even for languages with fewer speakers than English, using the Web as a last resort can help improve a system beyond what is possible using only a wordnet since both methods help for different kinds of relations (noun similarity and hyperonymy in the case of wordnets, versus instance relations for names in the case of Web search).

6.3.2 Semantic Classes

Non-overlapping semantic classes are a relatively cheap and robust way to incorporate semantic information into a coreference resolution system, and most systems, including but not limited to Soon *et al.* (2001) and Ng and Cardie (2002c), as well as Müller *et al.* (2002) for German (who use a hand-annotated *human/concrete/abstract* distinction) incorporate such a distinction.[16]

The approach I present in this section achieves acceptable classification results using a bare minimum of training data, as no annotated data is available for this purpose.[17]

[16] All systems that do coreference resolution on the ACE corpus incorporate a means of getting semantic classes according to the ACE scheme, as only members of these classes are to be linked in the coreference resolution task.

[17] The CoNLL shared task on named entities (Tjong Kim Sang and De Meulder, 2003) is limited

As development data for the creation of a classifier, a total of 936 non-singleton mentions was manually classified with a five-class inventory (PER, ORG, LOC, EVT, TMP, Obj, and Other, for persons, organizations, locations, events, temporal entities, concrete objects and everything else), and the following groups of features were implemented:

- One group of features looks up the synsets corresponding to the head lemma and searches for a set of pre-defined synsets in the hyperonym trees of the synsets, including `nMensch.1` (*Person*), `nMensch.574` (*Mensch_hierarchisch*), but also `nArtefakt.2719` (*Verkehrsweg*), `nGruppe.752` (*Organization*) and other upper-level nodes highly indicative of the semantic class.

- One group of features targeted more at named entities, checking for honorifics, organizational suffixes, and using gazetteer lists[18].

- Knowledge-poor patterns capturing morphological patterns such as three-letter acronyms (which often appear as organization names), binnen-I gender-neutral forms (as in *SchneiderInnen*), and a pattern to capture heads like *43-jähriger*.

Experiments were performed both using machine learning classifiers (see tables 6.5 through 6.8) and by modifying the rules from the J48 and JRip classifiers. While the classification accuracy is generally good (between 75.3%, or 119/158 examples from the test set, for the best-performing TiMBL classifier, and 77.8%, or 123/158 examples for the J48 decision tree), the precision for the *person* (PER) class is relatively low (67% to 74%, compared with 85%-90% for the other large classes of *organizations* and *locations*). Therefore, I decided to use a modified rule set where less precise rules are omitted and mapped to abstentions instead (see table 6.8).

The decision list proceeds as follows:

- if the mention contains a complete person name (from Biemann's `inDBperson` list), classify as `PER`.

- if the mention head matches the *...-jähriger* pattern, classify as `PER`.

- if the head is a hyponym of *Mensch_Beruf* in GermaNet, classify as `PER`.

- if the mention is a company name from the CDG lexicon, classify as `ORG`.

- if the mention head is a three-letter-acronym, classify as `ORG`.

to MUC named entities – names and name-related adjectives, whereas the Heidelberg Text Corpus used by Müller *et al.* (2002) contains a very different type of text.

[18]The gazetteer lists are derived from the WCDG parser's lexicon, the UN-ECE Locode database (`http://www.unece.org/cefact/locode/`), as well as a list of first and last names and person names compiled by Biemann (2002).

- if the mention has an organization suffix, classify as ORG.
- if the mention is a sequence of first and last name, classify as PER.
- if the mention is classified as Region in the CDG lexicon, classify as LOC.
- if the mention is a region in the UN-LOCODE gazetteer, classify as LOC.
- if the mention head is a hyponym of *Zeiteinheit* in GermaNet, classify as TMP.
- if the mention head can be a hyponym of *Ereignis*, but not of *Soziale_Gruppe* or *Bauwerk*, classify as EVT.
- if the mention head can be a hyponym of *Person*, but not of *Ereignis* or *Organization*, classify as PER.
- if the mention head can be a hyponym of *Organization*, but not of *Bauwerk*, *Ort* or *Artefakt*, classify as ORG.
- if the mention head's hyperonyms include all of *Ereignis*, *Gruppe* and *Organization*, classify as ORG.
- if the mention head can be a hyponym of *Gruppe*, but not of *Artefakt*, *Person*, *Ereignis* or *Ort*, classify as ORG.
- if the mention is in the location gazetteer list, classify as LOC.
- if the mention head can be a hyponym of *Verkehrsweg*, classify as LOC.
- if the mention head can be a hyponym of *Ort*, but not of *Ereignis*, *Person* or *Gruppe*, classify as LOC.

Table 6.11 compares the learned decision-tree and decision-list classifiers with the modified decision list in terms of accuracy, and their effect in the TUECOREF system that is presented in chapter 7. Despite the fact that the decision list does significantly worse in terms of overall accuracy, coreference results benefit from the increased precision of *person* classifications.

Improving sense distinctions Many of the common nouns that are to be classified are ambiguous between several semantic classes because they have multiple word senses in GermaNet. Several kinds of systematic polysemy are relatively frequent, including polysemy between persons and organizations (e.g. *Hersteller*, *Importeur*), organizations and their locations (e.g. *Altenheim*), or event-result polysemy (e.g. *Studie*). In many of these polysemous cases, one of the senses is clearly more frequent than the others - for example, *Hersteller* usually refers to companies in newspaper text.

	PER	ORG	LOC	EVT	TMP	Obj	Rest	*Prec*
PER	47	9	5	-	3	3	-	67.1
ORG	-	28	2	-	-	1	-	90.3
LOC	-	1	10	-	-	-	-	90.9
EVT	-	1	1	17	-	4	-	73.9
TMP	-	-	-	-	1	-	-	100
Obj	-	2	-	3	-	10	-	66.6
Rest	1	-	1	-	-	2	6	60.0
Recl	97.9	68.3	58.8	68.0	100	50.0	66.6	

Table 6.5: Confusion matrix for Timbl 5.1 classifier (`-mM -k7 -d ID`)

	PER	ORG	LOC	EVT	TMP	Obj	Rest	*Prec*
PER	46	5	2	4	-	3	2	74.2
ORG	1	28	2	-	-	1	-	87.5
LOC	-	2	12	-	-	-	-	85.7
EVT	-	3	1	20	-	6	-	66.7
TMP	-	-	-	1	1	-	-	50.0
Obj	-	2	-	-	-	8	-	80.0
Rest	1	1	-	-	-	2	7	63.6
Recl	95.8	68.3	70.6	80.0	100	40.0	77.8	

Table 6.6: Confusion matrix for WEKA JRip classifier (default settings)

	PER	ORG	LOC	EVT	TMP	Obj	Rest	*Prec*
PER	47	6	3	6	-	3	2	70.1
ORG	-	32	2	1	-	1	-	88.9
LOC	-	1	11	-	-	-	-	91.7
EVT	-	-	1	15	-	4	-	75.0
TMP	-	-	-	1	1	-	-	50.0
Obj	-	2	-	2	-	10	-	71.4
Rest	1	-	-	-	-	2	7	70.0
Recl	97.9	78.0	64.7	60.0	100	50.0	77.8	

Table 6.7: Confusion matrix for WEKA J48 classifier (default settings)

	PER	ORG	LOC	EVT	TMP	Obj	Rest*	*Prec*
PER	32	8	1	-	-	-	-	78.0
ORG	-	22	2	-	-	2	-	84.6
LOC	-	1	11	-	-	-	-	90.9
EVT	-	2	-	18	-	2	-	73.0
TMP	-	-	-	1	1	-	-	50.0
Obj	-	-	-	-	-	-	-	0
Rest*	16	8	3	6	-	13	9	16.9
Recl	66.6	53.5	55.5	95.0	7.1	0	100	

*: *Rest* includes abstentions of the classifier

Table 6.8: Confusion matrix for modified decision list
Rows: prediction of the classifier; Columns: semantic class according to gold standard

Mark		
nMenge.147	2.73	Deutsche Mark/DM/D-Mark
nPflanze.1735	1.20	Pflanzenteil/Zweig/Zapfenschuppe
nKoerper.341	1.20	Zahnmark/Rückenmark/Nierenmark
Organization		
nGruppe.753	5.00	politisches System/Wissenschaftsorganization
nGruppe.1214	2.98	Unterteilung/Systematisierung
nKognition.1598	1.70	Umorganization/Strukturierung
Importeur		
nGruppe.2369	1.92	Importunternehmen/Firma
nMensch.1331	1.33	Importkaufmann/Importhändlerin
Ausschuß		
nGruppe.2102	4.95	Kommission/Überwachungsausschuß
nArtefakt.3781	1.00	Nebenprodukt/Abfallprodukt

Table 6.9: Ranked word senses

	PER	ORG	LOC	EVT	TMP	Obj	Rest	*Prec*
PER	47	10	3	6	-	4	3	64.4
ORG	-	28	2	-	-	1	-	90.3
LOC	-	1	11	-	-	-	-	91.7
EVT	-	1	1	16	-	4	-	72.7
TMP	-	-	-	-	1	-	-	100
Obj	-	1	-	2	-	9	-	75.0
Rest	1	-	-	1	-	2	6	60.0
Recl	97.9	68.3	64.7	64.0	100	45.0	66.7	

Table 6.10: Confusion matrix for WEKA J48 classifier with most frequent sense information

	Classification	Coreference		
	Accuracy	Prec	Recl	F
J48	77.8	60.4	69.7	64.7
Jrip	77.2	60.5	69.5	64.7
mod-DL	58.9	60.3	70.7	65.0

Table 6.11: Influence of Semantic Tagging on Resolution Accuracy

While most-frequent-sense information is not available in GermaNet, McCarthy *et al.* (2004) present an approach to determine the most frequent sense using a distributional similarity model learned from large amounts of in-domain text. Using the distributional similarity model presented in section 6.2.5, it is therefore possible to such an approach for ranking senses according to estimated frequency.

McCarthy *et al.* exploit the observation that, in the presence of semantically ambiguous words, distributional thesauri usually have higher similarity scores for words semantically related to the most frequent sense (as the context distribution for the ambiguous word will approximate the hypothetical distribution of contexts that one would get by selecting only the most frequent sense).

To adapt this approach to German, a most similar word list is retrieved by decompounding the word with the method described in section 6.2.3, and retrieving the most similar items from the distributional similarity database, while synset lookup in GermaNet is done using the longest suffix matching technique used normally. This means that the forms used to look up the synsets and the most similar items can be different (for example, while *Zivilbevölkerung* is present both in GermaNet and the distributional thesaurus, *Einzeltäter* is found as such in GermaNet, while the distributional thesaurus backs this off to *Täter*, whereas for *Studentenfilmtag*, the entry *Filmtag* in the distributional thesaurus, but *Tag* in GermaNet would be used).

Given the fact that this approach does not use any other information, it does relatively well: *Mark* in the currency sense is preferred over *Mark* as marrow, and the organization/company sense of *Organization* is preferred over the event sense, and *Importeur* in the company sense is slightly preferred over the individual person sense (cf. table 6.9).

Unfortunately, adding most frequent sense information to the semantic tagger does not seem to help: with additional features indicating that the *highest-ranked* sense is a hyponym of the mentioned synsets, accuracy on the test set goes down (from 77.8% to 74.6%), as well as the precision for the person class; this is probably due to the fact that training and test set have a slightly differing composition.

A more determined approach to semantic classes – using a larger amount of data, and preventing skew in the semantic class distribution by also annotating singleton mentions – could conceivably raise the quality of the classification, and may make it possible to adjust the precision for selected classes by under- or oversampling. Considering the amount of training data, however, the semantic classification measurably helps the full coreference system in chapter 7 and provides a surprisingly strong baseline for the unsupervised, low-precision/high-recall approaches that are investigated in section 6.4.

6.4 Experiments on Antecedent Selection

Most approaches in the literature that aim at the evaluation of single features or small groups of features (Markert and Nissim, 2005; Garera and Yarowsky, 2006, *inter alia*) evaluate their model in the context of *antecedent selection*, where a discourse-old definite description is resolved to an antecedent. As less than half of all definite descriptions are discourse-old, such an evaluation is not realistic, but its simplicity allows to concentrate on finding an appropriate antecedent rather than tackling the complex interaction between discourse-new classification and antecedent selection.

In a setting similar to Markert and Nissim (2005), I evaluate the precision (proportion of correct cases in the resolved cases) and recall (correct cases to all cases) for the resolution of discourse-old definite noun phrases. Before trying to resolve coreferent bridging cases, the system looks for compatible antecedent candidates with the same lexical head and resolve to the nearest such candidate if there is one.

For the experiments described in this section, I used the first 125 articles of the coreferentially annotated TüBa-D/Z corpus of written newspaper text (Hinrichs *et al.*, 2005a), totalling 2239 sentences with 633 discourse-old definite descriptions, and the release 5 of GermaNet (Kunze and Lemnitzer, 2002), which is also the German-language part of EuroWordNet. Noun phrases with a demonstrative article (*dies*/*diese*/*dieser*) are included in the evaluation, but are not treated differently by the resolver.

Unlike Markert and Nissim, I did not limit the evaluation to discourse-old noun phrases where an antecedent can be found in the 4 preceding sentences, but also included cases where the antecedent is further away. As a real coreference resolution system would have to either resolve them correctly or leave them unresolved, I feel that this is less unrealistic and thus preferable even when it gives less optimistic evaluation results.

The precision is a mixture of the precision of the same-head resolver, usually above 80%, and the precision of the resolution for coreferent bridging, which is much lower than that for same-head cases. As a result, any improvement that results in the resolution of more coreferent bridging cases brings about a decrease in precision, with any gain in recall being accompanied by a similar loss in precision. This kind of evaluation behavior would not be helpful for understanding the performance of non-same-head resolution.

As an alternative, Markert and Nissim propose an evaluation method in which any anaphoric definite that is left unresolved is assigned the most recent antecedent, thereby forcing the resolution of all cases. However, Markert and Nissim's strategy does not fit the desideratum of measuring the precision of coreferent bridging resolution – using their evaluation method, any non-trivial method of resolving coreferent bridging will increase the precision, even if its own precision is only slightly better than a surface-distance baseline.

The solution I propose here is to include as an additional precision measure for coreferent bridging cases alone (i.e., number of correct coreferent bridging cases

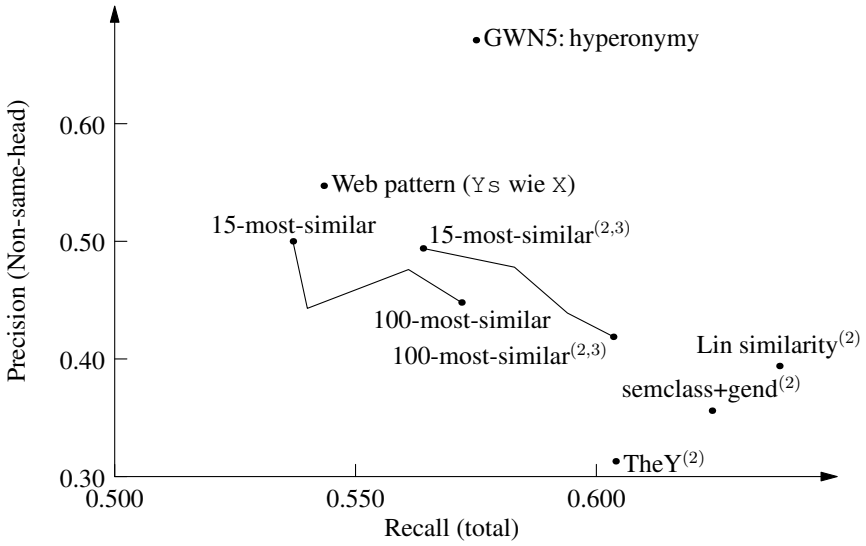
	Prec	Recl	$F_{\beta=1}$	Prec.NSH
same-head	0.87	0.50	0.63	—
nearest$^{(1)}$ (only number check)	0.57	0.55	0.56	0.12
semantic class+gender check$^{(1)}$	0.68	0.61	0.64	0.35
semantic class+gender check$^{(2)}$	0.67	0.62	0.65	0.36
GermaNet, hyperonymy lookup	**0.83**	0.58	**0.68**	**0.67**
GermaNet, node distance$^{(1)}$	0.71	0.61	0.65	0.39
single pattern: "Y wie X"$^{(1)}$	0.83	0.54	0.66	0.55
TheY$^{(1)}$ (only number checking)	0.66	0.59	0.62	0.29
TheY$^{(2)}$ (only number checking)	0.66	0.60	0.63	0.31
Lin$^{(1)}$ (only number checking)	0.66	0.60	0.63	0.30
Lin$^{(2)}$ (only number checking)	0.69	**0.64**	0.66	0.39
PL03$^{(1)}$ (only number checking)	0.68	0.63	0.65	0.38
PL03$^{(2)}$ (only number checking)	0.70	**0.64**	0.65	0.42
15-most-similar$^{(1)}$	0.82	0.54	0.65	0.50
100-most-similar$^{(2,3)}$	0.73	0.60	0.66	0.42

Prec.NSH: precision for coreferent bridging cases
$^{(1)}$: consider candidates in the 4 preceding sentences
$^{(2)}$: consider candidates in the 16 preceding sentences
$^{(3)}$: also try candidates such that the anaphor is in the antecedent's similarity list

Table 6.12: Baseline results

by all resolved coreferent bridging cases), which I take to correspond most closely to intuitive ideas of what 'better precision' would mean in an investigation of non-same-head resolution.

In my evaluation, I include hyperonymy search and a simple edge-based distance based on GermaNet, as well as a baseline using semantic classes (automatically determined by a combination of simple named entity classification and GermaNet subsumption), as well as the evolved version of Markert and Nissim's approach presented in section 6.1.2. For the methods based on similarity and association measures, I include a simple ranking by the respective similarity or relatedness value. Additionally, I include an approach due to Gasperin and Vieira (2004), who tackle the problem of similarity by using lists of most similar words to a certain word, based on a similarity measure closely related to Lin's. They allow resolution if either (i) the candidate is among the words most similar to the anaphor, (ii) the anaphor is among the words most similar to the candidate, (iii) the similarity lists of anaphor and candidate share a common item. I tried out several variations in the length of the similar words list (Gasperin and Vieira used 15, I also tried lists with 25, 50 and 100 items). The third possibility that Gasperin and Vieira mention (a common item in the similarity lists of both anaphor and antecedent) resolves some correct cases, but leads to a much larger number of false positives, which is why it is not included in the evaluation.



(2): consider candidates in the 16 preceding sentences
(3): also try candidates such that the anaphor is in the anteceden't similarity list.

Figure 6.5: Baseline results (graphic)

As can be seen in table 6.12, same-head resolution (including a check for modifier compatibility) allows to correctly resolve 49.8% of all cases, with a precision of 86.5%. The most simple approach for coreferent bridging, just resolving coreferent bridging cases to the nearest possible antecedent (only checking for number agreement), yields very poor precision (12% for the coreferent bridging cases), and as a result, the recall gain is very limited. Using semantic classes (based on both GermaNet and a simple classification for named entities) to constrain the candidates and then use the nearest number- and gender-compatible antecedent[19], we get a much better precision (35% for coreferent bridging cases), and a much better recall of 61.1%. Hyponymy lookup in GermaNet, without a limit on sentence distance, achieves a recall of 57.5% (with a precision of 67% for the resolved coreferent bridging cases), whereas using the best single pattern (Y wie X, which corresponds to the English Ys such as X), with a distance limit of 4 sentences[20], on the Web only improves the recall to 54.3% (with a lower precision of 55% for coreferent bridging cases). This is in contrast to the results of Markert and Nissim, who found that Web pattern search performs better than wordnet lookup; see (Versley, 2007) for a discussion. Ranking all candidates that are within a distance of 4 hyper-/hyponymy edges in GermaNet by their edge distance, I get a relatively good recall of 60.5%, but the precision (for the coreferent bridging cases) is only at 39%, which is quite poor in comparison.

[19] In German, grammatical gender is not as predictive as in English as it does not reproduce ontological distinctions. For persons, grammatical and natural gender almost always coincide, and I check gender equality whenever the mention to be resolved denotes a person.

[20] There is a degradation in precision for the pattern-based approach, but not for the GermaNet-based approach, which is why I do not use a distance limit for the GermaNet-based approach.

	Prec	Recl	$F_{\beta=1}$	Prec.NSH
sem. class+gender checking	0.68	0.61	0.64	0.35
GermaNet, hypernymy lookup	0.83	0.57	0.68	0.67
GermaNet < "Y wie X"	0.81	0.60	0.69	0.63
GermaNet < all patterns	0.81	0.61	0.70	0.64
TheY$^{(2)}$+semclass+gender	0.76	0.60	0.67	0.47
TheY+sem+gend+Bnd	0.78	0.59	0.67	0.50
Lin$^{(2)}$+semclass+gender	0.71	0.63	0.67	0.43
Lin+sem+gend+Bnd	0.80	0.58	0.67	0.53
PL03$^{(2)}$+semclass+gender	0.72	0.64	0.68	0.45
PL03+sem+gend+Bnd	0.80	0.59	0.68	0.57
GermaNet < all patterns	0.81	0.62	0.70	0.64
< 25-most-similar$^{(2,3)}$	0.79	0.65	0.72	0.62
< LinBnd	0.79	0.68	**0.73**	0.63
< Lin < TheY+sem+gend	0.74	**0.70**	0.72	0.54

$^{(2)}$: consider candidates in the 16 preceding sentences
$^{(3)}$: also try candidates such that the anaphor is in the antecedent's similarity list

Table 6.13: Combination-based approaches

The results for Garera and Yarowsky's TheY algorithm are quite disconcerting – recall and the precision on coreferent bridging cases are lower than the respective baseline using (wordnet-based) semantic class information or Padó and Lapata's association measure. The technique based on Lin's similarity measure does outperform the baseline, but still suffers from bad precision, along with Padó and Lapata's association measure. In other words, the similarity and association measures seem to be too noisy to be used directly for ranking antecedents. The approach of Gasperin and Vieira performs comparably to the approach using Web-based pattern search (although the precision is poorer than for the best-performing pattern for German, "X wie Y" – X such as Y, it is comparable to that of other patterns).

Combination-based approaches The above information sources draw from different kinds of evidence and thus should be rather complementary. In other words, it should be possible to get the best from all methods, achieving the recall of the high-recall methods (such as using semantic class information, or similarity and association measures), with a precision closer to the most precise method using GermaNet.

As a first step to improve the precision of the corpus-based approaches, I added filtering based on automatically assigned semantic classes (persons, organizations, events, other countable objects, and everything else). Very surprisingly, Garera and Yarowsky's TheY approach, despite starting out at a lower precision (31%, against 39% for Lin and 42% for PL03), profits much more from the semantic filter and

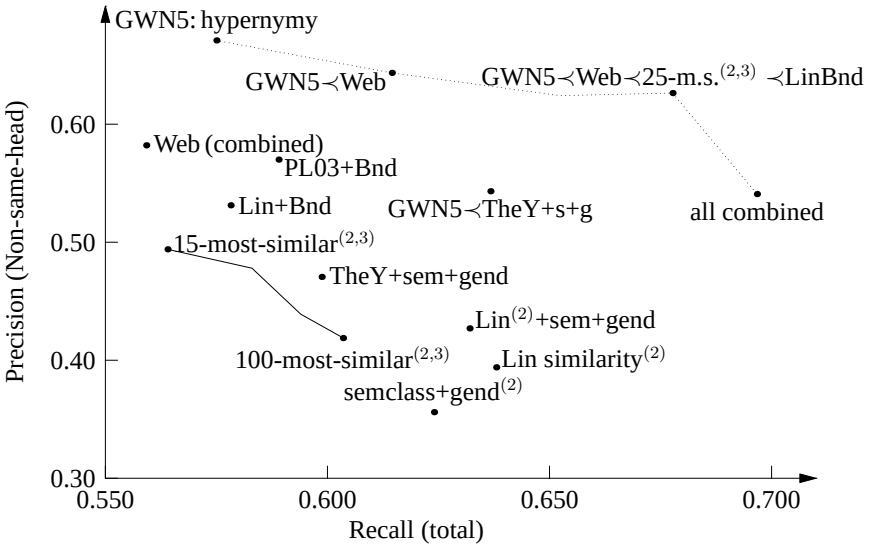


Figure 6.6: Combination-based approaches (graphic)

reaches the best precision (47%), whereas Lin's semantic similarity measure profits the least.

Since limiting the distance to the 4 previous sentences had quite a devastating effect for the approach based on Lin's similarity measure (which achieves 39% precision when all the candidates are available and 30% precision if it choses the most semantically similar out of the candidates that are in the last 4 sentences), I investigated a way to apply the distance-based filtering after finding semantically related candidates.

The approach I tried was as follows: the system ranks all candidates using the similarity function, and keep only the 3 top-rated candidates. From these 3 top-rated candidates, only those within the last 4 sentences are kept. Without filtering by semantic class, this improves the precision to 41% (from 30% for limiting the distance beforehand, or 39% without limiting the distance). Adding filtering based on semantic classes to this (only keeping those from the 3 top-rated candidates which have a compatible semantic class and are within the last 4 sentences), we get a much better precision of 53%, with a recall that can still be considered good (57.8%). In comparison with the similarity-list-based approach, the distance-bounding approach results in a much better precision than we could be achieved with methods having comparable recall (the version with the 100 most similar items has 44% precision, the version with 50 most similar items and matching both ways has 46% precision).

Applying this distance-bounding method to Garera and Yarowsky's association measure still leads to an improvement over only semantic class and gender compatibility, but the improvement (from 47% to 50%) is not as large as with the semantic similarity measure or Padó and Lapata's association measure (from 45% to 57%).

For the final system, I back off from the most precise information sources to the less precise. Starting with the combination of GermaNet and pattern-based search on the World Wide Web, we begin by adding the distance-bounded semantic similarity-based resolver (LinBnd) and resolution based on the list of 25 most similar words (following the approach of Gasperin and Vieira 2004). This results in visibly improved recall (from 62% to 68%), while the precision for coreferent bridging cases does not suffer much. Adding resolution based on Lin's semantic similarity measure and Garera and Yarowsky's TheY value leads to a further improvement in recall to 69.7%, but also leads to a larger loss in precision.

6.5 Summary

In this chapter, I have presented a variety of approaches for semantic information sources that can inform the resolution of definite description, together with implementations for German, accounting for issues such as synthetic compounds and morphological flexibility.

The approaches have been evaluated in the task of selecting an antecedent for a discourse-old definite description. The antecedent selection task is well-established in the literature but should not be confused with coreference resolution of definite descriptions as the (complete) coreference resolution task involves interaction between the discrimination of discourse-new and discourse-old mentions and the actual selection of antecedents. In contrast, antecedent selection is simpler and (as we will see) more friendly towards recall-oriented approaches. A more comprehensive system to tackle the actual coreference resolution of definite descriptions and names will be presented in chapter 7.

In a novel approach to combining distributional semantic measures with a filter based on accurate semantic classification, I have shown that both relation-based and relation-free similarity measures can be used to find plausible antecedents for anaphoric definite descriptions. In comparison to solutions that were presented in the literature as means to solve the problem of recall-oriented antecedent selection (Markert and Nissim, 2005; Garera and Yarowsky, 2006), both a simpler approach of ranking by a single similarity measure and the approach of using a similarity measure in conjunction with a semantic class-based filter achieve greater recall.

With 70% F-measure for the best approach that combines all information sources (GermaNet, Web patterns, distributional similarity), the outlook for resolution of discourse-old definite descriptions is quite good, but it has to be pointed out that the high-recall methods presented in this chapter are not precise enough as an indicator for whether a definite description has an antecedent (and, by consequent, needs to be resolved) or not.

One avenue for future work that has not been explored here would be ways to improve the approach of cross-sentence association by using the large-coverage information sources presented here; previous approaches to unsupervised or semi-supervised coreference resolution either did not factor in semantics at all (Müller *et al.*, 2002) or, as Haghigi and Klein (2007), made unrealistic assumptions about the unlabeled training data that limited the 'unlabeled' data to annotated coreference corpora.

Chapter 7

Description of the TUECOREF System

This chapter presents in greater detail a system for coreference resolution that identifies mentions in treebank trees and subsequently carries out a combination of anaphoricity detection and resolution for definite noun phrases for German on the TüBa-D/Z corpus. In contrast to the experiments in section 6.4, the experiments in this chapter aim at a more realistic setting where mentions are not known a priori to be discourse-old, and an approach to integrate discourse-old/discourse-new classification is needed instead.

The system presented here uses a novel approach to integrate the choice of resolution or non-resolution with the choice of possible antecedents using a maximum-entropy ranking resolver. Several of the information sources presented in chapters 5 and 6 are used as features to allow the resolution of coreferent bridging (i.e. cases where the antecedent of an anaphoric definite description has a different head). In contrast to the experiments in chapter 6, the TUECOREF system covers not only definite descriptions but also names.

Section 7.1 discusses the task definition used and the general system architecture. Section 7.1.3 provides details on the non-machine-learning components of the system, including the extraction of markables and hard constraints on same-head resolution. Section 7.1.4 presents the ranking classifier, and section 7.2 presents details on the approach to resolve coreferent bridging.

7.1 Coreference Resolution Framework

The TUECOREF system was developed using the referentially annotated TüBa-D/Z corpus as a source of annotated data. I used articles 1–125 for development (about 39 000 tokens, or 2240 sentences, with 3678 full noun phrases containing a definite or demonstrative article and 1766 names) and 126–250 for training (about 52 000 tokens, or 2898 sentences, with 4736 full noun phrases containing a definite or demonstrative article, and 2844 names).

The data set chosen is large enough to validate the constraints and the information from unsupervised learning that was used; as some of the features used in the system are relatively expensive to compute (especially the pattern search on the Web, where every pair of mentions to be tested results in four search engine queries), using a smaller training set is more practical than training on all the data available.

In a competitive setting, it would be advisable to use a *blind test set* which is standardized and is not used for system development so that a fair comparison of different systems is possible. Given that the focus of this work is on assessing the utility of different information sources - necessarily so, since the first results for resolving definite descriptions on the TüBa-D/Z were reported in (Versley, 2006), and subsequent work by Klenner and Ailloud (2008) or Broscheit *et al.* (2010) does not provide a separate evaluation for non-pronouns - the utility of one additional evaluation figure for blind testing would be relatively limited.[1]

The amount of data chosen for training and development set is still larger than the corpora chosen by Markert and Nissim, 2005 (who used 565 anaphoric definite noun phrases with an antecedent within current and last 4 sentences, extracted from the MUC-6 corpus), or the data used for classification by Garera and Yarowsky, 2006 (who used 177 hand-selected definite descriptions from the English Gigaword Corpus, covering 13 distinct hyperonyms).

The selection of texts is also larger than the corpora used by Hartrumpf (2001), who selected 12 texts containing 502 discourse-old noun phrases (pronouns, definite descriptions, names), or Müller *et al.* (2002), whose corpus comprises 2179 discourse-old noun phrases (pronouns, definite descriptions, names) in about 37 000 tokens. (The training set of TüBa-D/Z files 126-250 contains 1556 discourse-old definite descriptions and names).

Evaluation is carried out in the *link-based* evaluation scheme described in section 3.2.1.

7.1.1 Mention Identification

Coreference resolution as presented in this thesis concerns linking potentially reference-bearing linguistic entities – in this case, full noun phrases, including definite and indefinite common noun phrases as well as names.

However, there is no one-to-one correspondence between NX nodes in the TüBa-D/Z treebank and referential indices:[2] The TüBa-D/Z uses an adjunction-

[1] Doing a comprehensive assessment of different features and/or system settings as it is done for the development set would defeat the purpose of a blind test set if it were done there as it would effectively make the test set non-blind. Note that no automatic feature selection or tuning of (hyper-) parameters was used for the experiments reported here. Such techniques would have to be carried out on a jackknifed portion of the training set rather than the development set, as they lead to considerable overfitting to the data set used for tuning or parameter selection.

[2] Karttunen (1976) uses the term *referential index* citing Chomsky's (1965) use of the term as an abstract notion for referential identity that is used for syntactic constraints such as the *i*-within-*i* criterion. For the presentation here, it will serve as a theory-neutral term for "things we stick a $[\,]_i$ on

like scheme for postmodification of NPs; appositions, which consist of multiple noun phrases nodes, only correspond to one single referential index; and some noun phrases (e.g., in copula constructions) cannot introduce a referent by themselves.

Therefore, a necessary preliminary step in coreference resolution is the identification of the NP projections corresponding to one referential index (i.e., mention). In most cases, this referential index is the projection of one head noun. From the phrase structure tree, all phrasal projections are retrieved, so that coreference set IDs from the gold annotation can be related to the mention objects introduced.[3]

The collection of noun phrase projection also serves to prevent inconsistencies due to erroneous annotations: In the case of "*the man with the telescope*", the whole NP is to be marked as a coreference markable, but in some cases annotators only mark a smaller span, such as "[*the man*] *with the telescope*". Training and evaluation treat a mention as a correct antecedent whenever any of the projections has been annotated as the target of a coreference link (or in the set-based annotation, a member of the same coreference set).

A mention object is created for every token belonging to a predefined set of parts of speech (for the "coref" class, these are `NN` and `NE`), unless they are in one of the following positions:

- any token in a name or apposition except the first (corresponding to the `APP` dependency label in Foth's scheme). This also encompasses cases of postmodification such as "AWO *Bremen*", which do not introduce a reference to *Bremen*.
- items occurring as predicate in copula constructions (John is *a farmer* – `PRED` dependency label)
- noun phrases occurring with *als*: Peter arbeitet *als Bauarbeiter* (*Peter works as a construction worker*
- Vorfeld-es and correlates: Ich finde *es* schade, dass nichts passiert (*I consider it a pity that nothing happens*, `EXPL` dependency label).[4]

It would be desirable to introduce mention objects for coordinated noun phrases, as these introduce a referential index, and also to distinguish between role (als/*as*) and comparative (als/*than*) readings of *als*. However, for the sake of simplicity, this has been omitted.

The mention objects corresponding to the referential indices are then enriched with morphological, syntactic and semantic properties that can subsequently be used in filtering out incompatible antecedents or in features for ranking.

---

and which play a role in the establishment of discourse referents in the syntax-semantics interface".

[3] In the more recent PALinkA-based annotations, the link-based annotation is converted to a set-based annotation for this purpose.

[4] As pronouns are generally not considered plausible antecedents, this does not make any difference for the definite description resolution in the current TUECOREF system.

Based on determiners and determiner-like premodification such as premodifying genitives and non-determiner attributive indefinite pronouns (PIDAT; e.g., *all die Blumen*, *die vielen Leute*) the mention is assigned a determination type:

- **def** for definite articles (*der*, *die*)
- **indef** for indefinite articles (*ein*, *eine*)
- **poss** for genitive premodification ("*Peters* Tagebuch") or attributive possessive pronouns ("*sein* Tagebuch")
- **quant** for indefinite attributive pronouns (*solche*, *viele*, *keine*, *mehr*), as well as NPs with premodifying cardinal numbers ("*86* Millionen Mark").

Only name markables (which contain a headword tagged as NE) and those with a *def, poss*, or *pdat* determination type are included in the resolution process. Other mentions with a common noun head and indefinite or quantifying determination are counted as a false negative when they are discourse-old according to the annotation, but are never resolved by the system. Foreign-word mentions (such as "*Hajusom!*" or "*Bodysnatchers*", which occur five times in the training set and once in the development set) are excluded from both resolution and scoring.

The `mods` slot in the mention is used to record premodifying adjectives (`ATTR`), as well as post-modification by genitives (`GMOD`), prepositional phrases (`PP`), relative phrases (`REL`), or modifying infinitive clauses (`OBJI`).

Postmodifying region identifiers, such as *Bremen* in *die Arbeiterwohlfahrt Bremen* should not be matched to the actual region (in the example *Bremen*), which is why they are removed from the list of name parts and instead added to the modifier list with a `COMP` label.

7.1.2 System Architecture

The basic building block of the system is the *generate-filter-rank* approach: for each mention, a **generator** proposes a set of potential antecedents. In a second step, these antecedents are then filtered by a set of **hard (filtering) constraints**, and in the last step the remaining antecedents are ranked according to a set of **soft (ranking) constraints**.

Same-head resolution and coreferent bridging are realized as two separate resolvers that are invoked in sequence:

- First, the system tries to find a *same-head antecedent* through a generator that proposes same-head antecedents and various filters that check linguistic compatibility. Unless otherwise stated, the system chooses between same-head antecedents by picking the closest one.
- If no antecedent is proposed by the same-head component, several strong indicators for a *unique* description are checked. Descriptions which correspond to one of these uniqueness indicators (for example, all names) are

treated as discourse-new and not considered further. More detail on the hard uniqueness indicators is provided below in section 7.2.

- For *coreferent bridging* resolution, the system looks for potential antecedents with any head, which however have to be relatively close to the definite description. These potential antecedents are, again, filtered to check linguistic compatibility. A maximum entropy ranker then selects either one of the potential antecedents or an additional item that marks the definite description under consideration as discourse-new.

The system never undoes any of its resolution decisions, unlike the systems of Luo *et al.* (2004) or Daumé III and Marcu (2005) (see the description in section 4.2.2), which do a beam search over all possible coreference chains and are able to use constraints over whole coreference chains instead of simply considering each pair in isolation. In some rare cases, backtracking helps avoid errors: once we linked a mention *Clinton* to a previous mention of *Mr. Clinton*, we should not use the same mention *Clinton* as the antecedent of a mention such as *Hillary Clinton* or *she*.

As the focus of the thesis is the integration of features expressing semantic compatibility, I decided to avoid the additional complexity that is needed to keep track of multiple alternative sets of coreference chains. Without non-local features – the focus of the TUECOREF system is on providing good information on the compatibility between a definite description and a potential antecedent – the top-ranked choice is always the (globally) optimal solution, and backtracking is never needed.

7.1.3 General setting and Same-head resolution

The resolution to same-head antecedents largely hinges on the use of appropriate filtering constraints, which do not modify the distance-based ranking in any way (except by changing the input to the ranking). To quantify the impact that resolution heuristics have on the system and thereby aid the development of linguistically sensible heuristics, a glass-box approach to evaluation is very appropriate since it allows to compare expected and actual effect of an added constraint while mitigating the risk of overfitting the training/development data that would be present in blindly applying those changes that increase an overall evaluation figure.

For the same-head resolution component, I give *upper and lower bounds* on precision and recall for each variant I consider, based on the candidate sets in the training corpus, after filtering with the hard constraints. The upper bounds on recall and precision attainable through the ranking performed by the soft constraints are denoted by R_{max}/P_{max} in the table, and the lower bounds on recall and precision by R_{min} and P_{min}.

See table 7.1 for different configurations and the resulting precision and recall bounds and actual values for precision and recall on the development set. The table also contains a column with a *perplexity* value, which describes how easy it is for

	Pmax	Rmax	Pmin	Rmin	Perp	Prec	Recl	F
baselines								
all	27.0	98.7	0.0	0.0	23.68	1.2	4.9	1.9
4gram	31.1	76.6	13.3	37.5	2.28	26.3	**54.7**	35.5
head identity	52.1	54.4	32.1	47.1	1.68	58.2	50.5	54.1
J48 decision tree	NA	NA	NA	NA	NA	**76.0**	45.1	**56.6**
force resolution								
same_head	49.0	76.9	33.6	59.0	1.65	51.6	69.4	59.2
+agr_num	52.1	76.5	36.3	60.4	1.62	56.0	**69.7**	62.1
+comp_mod	56.4	71.4	38.2	57.7	1.57	62.1	64.8	63.4
uniq_name	57.1	74.3	40.5	61.6	1.57	62.0	68.6	65.1
+hard_seg(8)	64.9	68.7	43.8	59.0	1.61	**67.8**	63.2	65.4
+loose_seg(8)	62.8	71.1	43.0	59.8	1.58	66.6	65.8	**66.2**
always allow non-resolution								
head identity	100.0	54.4	0.0	0.0	1.89	62.5	38.5	47.6
same head	100.0	76.9	0.0	0.0	1.98	58.3	40.5	47.8
uniq_name	100.0	74.3	0.0	0.0	1.88	**66.8**	**58.4**	**62.3**

same_head: lemmatization+suffix matching
agr_num: check number agreement
comp_mod: check modifier compatibility
uniq_name: require matching name when matching NEs
unique_mod: check for modifiers that are typical for unique definites
segment: hard limit on sentence distance (4 sent.)

Actual precision (Prec,Recl) has been determined on the evaluation dataset; upper/lower bounds on precision and recall (Pmax,Pmin,Rmax,Rmin) were determined on the training dataset.

Table 7.1: Upper and lower bounds / same-head resolution

the ranker to chose the correct alternative among the candidates that are left after the hard filtering stage.

The perplexity is an information-theoretic measure that reflects how difficult it is, on average, to make the correct choice in the cases where multiple alternatives are left. Assuming that the ranker assigns a probability distribution $\hat{P}$, the perplexity is inverse of the (geometric) average of the likelihood of the correct choice given the alternatives and the (soft) ranking constraints (i.e., $\frac{1}{\hat{P}(y|x)}$).

In the case where the soft (ranking) constraints do not provide any information at all, the perplexity is equal to (the geometric average of) the number of candidates a system has to choose from. More information from the soft (ranking) constraints reduces this number; if the information always allows a decision with total certainty (i.e., a probability of 1.0), the perplexity becomes 1.0.

For the case of same-head resolution, the only soft (ranking) constraint is a *distance* constraint which ranks closer sentences higher than more distant ones. This corresponds to an exponential distribution: an antecedent candidate y_1 that is k sentences further back than another candidate y_2 is judged to be less likely by a factor of α^k (where α is chosen so that the perplexity is minimal).

Baseline results The first part of table 7.1 contains baseline results, with two relatively loose criteria (resolving to *any* previous noun phrase or resolving to any previous noun phrase with a minimum of surface similarity), one slightly stronger, but still shallow criterion, and a decision-tree-based resolver that serves as a stronger *intelligent baseline*.

The first result – resolving all definite noun phrases – yields precision and recall bounds that are instructive with respect to the task at hand: The upper recall bound is at 98.7%, which means that 1.3% of discourse-old full noun-phrases have been introduced by a cataphoric pronoun (and cannot be correctly resolved to a non-pronoun antecedent). The upper precision bound in this case (27.0%) corresponds to the proportion of definite full noun phrases that are discourse-old – the rest, about 73%, are discourse-new.

The baseline named *4gram* aims towards setting an upper baseline for the proportion of candidates that can be resolved using methods that are purely based on the surface strings of the mentions. Mentions are often varied in a text – for example *Ute Wedemeier* being mentioned again as *Frau Wedemeier*, and morphological variation in the case of common nouns, for example *die Angestellte* and *der Angestellten* – so that string equality is no longer sufficient to indicate a surface match, even though the head noun is basically the same. In the case of compounds, such as *Haus* and *Wohnhaus*, the strings are not identical, yet it should be possible to identify such cases using only linguistic knowledge.

All these cases have in common that a substantial amount of the surface string is shared. Strube *et al.* (2002) propose to use edit distance as a knowledge-poor method to achieve useful recall on such cases where strings are not identical.

In the remainder of this section, I will demonstrate that it is possible to achieve an optimal recall bound on surface-similar cases, at improved precision, using linguistic knowledge.

For the 4gram baseline, any candidate that shares a simple letter 4gram with the (potential) anaphor is seen to be valid – for example, *Haus* and *Wohnhaus* share the letter 4gram "*haus*", and *Angestellte* and *Angestellten* share a substantial number of letter 4grams. Of course, looking at shared 4-grams also yields a number of false positives, such as *Feuersbrunst* and *Aktionärsbrief* sharing the 4-gram *rsbr*. The goal for same-head resolution is to achieve the upper recall bound of 76.6% that we get with the 4gram technique, yet with a precision that is much better than the 31.1% upper bound that we can achieve using only this very shallow technique.

For comparison, requiring that one of the head nouns (for "*Frau Wedemeier*", both *Frau* and *Wedemeier* are considered heads) is exactly identical yields a much decreased recall bound of 54.4% (see table 7.1 under *head identity*), but a better precision bound of 52.1%.

The decision-tree baseline As an intelligent baseline, I reimplemented the decision-tree-based approach of Strube *et al.* (2002), with the feature set presented in their paper using my automatically derived, but more detailed, semantic classes and Weka's[5] J48 decision tree classifier. Because the dataset treats appositional constructs (as in *Peter, the CEO*) as one single markable, I did not use the *apposition* feature mentioned in their paper. For the surface-based similarity features (string identity, substring match and the edit distance-based features), I treated two mentions as compatible if any appositional part was (i.e., *Peter, the CEO* would be similar both to *Peter* and to *the CEO*). My implementation yields results of 76.0% precision and 45.1% recall (F=26.4% for definite descriptions and F=80.7% for names).

These figures are not comparable to those of Strube *et al.*, since the text type is different (touristic brochures versus newspaper text) and the annotation schemes contain significant differences (Strube *et al.* treat apposition as part of the coreference task, whereas the TüBa-D/Z annotation scheme assumes them as part of the preprocessing, since they are part of the syntax annotation). Nonetheless, the results are similar enough to the best results reported by Strube *et al.* (who report F=33.9% for definite descriptions and F=76.2% for names) to establish the decision tree system as a strong baseline.

Head matching with Linguistic Constraints As mentioned before, German synthetic compounds and richer morphology make it impossible to achieve good performance with string matching only (in contrast to English, where simpler morphology and analytic compounds mean that just comparing strings yields useful results). Therefore, it is necessary to take into account these factors to avoid the large loss in recall that we see with the *head identity* baseline.

[5] Weka is a Java machine learning library; `http://www.cs.waikato.ac.nz/ml/weka`

Taking into account the problems mentioned above, the *same head* resolver marks a mention as a possible antecedent if:

- two head strings are identical (e.g., "*Ute Wedemeier*", "*Frau Wedemeier*")
- one of the head lemmas for the candidate has one of the lemmas for the mention to be resolved as its suffix (e.g., "*Wohnhaus*", "*Haus*") or
- the surface strings of either head are within one edit operation of each other and neither is shorter than four letters (e.g. "*Bremens*" and "*Bremen*").

Lemmatization is done using SMOR (Schmid *et al.*, 2004) and the lexicon of the WCDG parser (Foth *et al.*, 2004). In cases where a lemma is ambiguous, all possible lemmas are considered for matching.

As indicated in table 7.1, this approach yields an upper recall bound of 76.9%, which is even slightly better than the upper bound that results from 4-gram matching, but with a much improved precision bound of 49.0%.

The low precision – even the *head identity* baseline has a relatively low precision bound of 52.1% – is due to spurious antecedents found for mentions which are either discourse-new (and should not be resolved to a prior mention at all), or for mentions which have a non-same-head antecedent (which is not included in the possible antecedents for same-head resolution).

While some false positives cannot be avoided in the case of generic mentions (where no reference to a specific entity is made, and coreference should not be annotated following the annotation guidelines), there are some spurious matches due to cases where two mentions share the head noun, but do not corefer.

Checking *number agreement* solves some of these cases, improving the upper precision bound to that for the 'head identity' variant, without any noticeable impact on recall.

Another source for spurious matches are pairs of mentions such as *the red car–the blue car*, where two instances of a single concept are mentioned, but are incompatible due to having different modifiers.

I use heuristics similar to those of Vieira and Poesio (2000) to filter out such cases, notably requiring that all modifiers present in the anaphor are also present in the antecedent. An exception is made for adjectives, where it is allowed that the anaphor has some attributive adjectives when the antecedent doesn't have any at all. (In the case of an anaphoric definite NP such as "*the red car*", "*a car*" can be a plausible antecedent, but "*a blue car*" would never be). The result of checking modifier compatibility is a large increase in both the upper bound on precision (from 52.1% to 56.4%) and in the actual precision on the development set (from 56.0% to 62.1%). However, some positive examples are filtered out by this treatment and recall is affected – the recall bound drops from 76.5% to 71.4%, and actual recall on the development set drops from 69.7% to 64.8%.

Treatment of names All heuristics up to here – the *same head* criterion, as well as modifier compatibility, have treated names and common noun phrases alike.

But named entities are special in that names are usually unique to an entity (e.g. "*Miller, the CEO*" is different from "*Smith, the CEO*", although they both share the common noun *CEO*). Conversely, different modifiers are not always a sign of incompatibility since names specify uniquely. (For example, "*John Miller*" and "*the overly clever Mr. Miller*" are perfectly compatible).

Therefore, two named entities are only allowed to match if they share the name, not if they share any other common noun. Named entities also occur more frequently in conjunction with modifiers that are indicative of uniqueness, even when they are discourse-old (for example *the more expensive ID Bremen*), which is why I do not check for modifier compatibility in the case of named entities.

Location postmodifiers (such as *Bremen* in *ID Bremen*) frequently lead to false positives despite having the same label as head nouns in the dependency representation (APP). Therefore, such postmodifiers are heuristically identified by looking for location names at the end of the head words list – if a noun phrase has several words in its head word list and it is not headed by a common noun such as *Stadt* (city) – and such location names are then excluded from the head matching.

The net result of this improved treatment of names (called *uniq_name* in table 7.1 is a small improvement in the upper and lower bounds relative to the *same head* version with number agreement and modifier compatibility – lower bounds on precision and recall improve quite visibly, from 38.2% to 40.5%, (an absolute gain of 2.3%) for precision, and from 57.7% to 61.6% for recall (an absolute gain of 3.9%). Actual recall on the evaluation data set is also improved, from 64.8% to 68.6%, while actual precision is unaffected. Comparing to the earlier *same head* result with number agreement but no modifier compatibility, precision has improved by a large amount – from 56.0% to 62.0%, or a 6% absolute gain, whereas the loss in recall – from 69.7% to 68.6%, or a 1.1% absolute loss – is relatively modest. Both precision and recall on the development set are beyond the *4gram* and *head identity* baselines, which suggests that the linguistically informed treatment of surface-similar antecedent candidates proposed here is successful.

Segmentation heuristics In the *force resolution* group of table 7.1, resolution is always attempted if a same-head antecedent is found, regardless of the distance. A *segmentation* heuristic would limit the distance that a potential antecedent can be away. The fact that antecedent candidates that pass the hard filter are only ranked by distance means that the choice of antecedents can not be influenced, but only whether the system chooses an antecedent or not.

In the section of table 7.1 named *always allow non-resolution*, the ranker described in section 7.1.4 is used to rank among the antecedent candidates plus an additional *do not resolve* pseudo-candidate; as distance is still the only ranking constraint among the antecedent candidates, the system can still only resolve to the closest of the antecedent candidates. However, the weight for the do-not-resolve

pseudo-candidate results in a distance limit beyond which the candidate is always ranked below the do-not-resolve pseudo-candidate. In this fashion, precision can be improved by a considerable amount, but recall suffers accordingly.

Similarly, imposing a hard window of eight sentences back (i.e., refusing to resolve to an antecedent candidate that is more than eight sentences away) improves the precision by more than 5% over the *unique names* case, but has a detrimental effect of the same size on recall.

Vieira and Poesio (2000) introduce a loose segmentation heuristic where they consider antecedents that are either not further away than a certain number of sentences or have been mentioned multiple times. In my case, the loose segmentation heuristic improved precision by the a similar amount to hard segmentation, but with a much smaller loss of recall. As the loose segmentation relies on information about the (potential) antecedent being an initial or subsequent mention, the actual results (66.6% precision and 65.8% recall, see the "*loose_seg*" column in table 7.1) are below what would be possible with perfect information on the information status of mentions (which would yield 67.8% precision and 67.0% recall).

Comparison to the decision-tree baseline For the case of same-head resolution, two variants of the system achieve a good balance of precision and recall: one is the *unique names* variant, which achieves 62% recall and 68.6% precision, or F=65.1%, and the other is the variant with Vieira and Poesio's *loose segmentation* heuristic added to the system, which achieves more balanced values of 66.6% precision and 65.8% recall (or 66.2% F-measure).

The decision-tree based system achieves much higher precision (76.0%) at a large cost of recall (45.1%) – in consequence, the more balanced approaches have a drastically higher F-measure (66.2% vs. 56.6%). However, it may be useful to have this kind of relatively high precision in a system, and a closer look is warranted to see exactly *how* this high precision is achieved.

For names, the decision tree system has 77.5% recall and 83.2% precision, yielding F=80.2%, whereas my system has 90.0% recall at 84.4% precision, yielding F=87.5%. For definite descriptions (non-names), my system has a recall of 50.7% and a precision of 43.6%, whereas the decision tree system has only 16.4% recall, at 55.4% precision.

In other words, much of the recall for definite descriptions is sacrificed for a gain in precision that seems small in relation. Because the end result is an average of name and definite description resolution, the higher-precision name resolution dominates the result for the decision tree system, yielding a higher precision.

If one wanted to achieve a higher precision using the *unique name* system as a starting point, one could simply remove *all* definite descriptions from the resolution, yielding performance figures of R=42.4%, P=84.4% (instead of R=45.1%, P=76.0% for the decision tree system), which looks even more appealing for applications where high precision is needed. (But see the section on p. 234 for a variant that uses the semantic classes to achieve more recall at a similarly high precision).

7.1.4 Learning a ranking resolver

The same-head resolver presented in the previous parts of this section shows that it is possible to achieve high performance on the same-head cases by using hard (filtering) constraints that exploit linguistic regularities in the pairing of a potentially anaphoric noun phrase and an antecedent candidate.

For the case of non-same head resolution, it is necessary to combine a larger number of semantic and salience indicators, allowing for a selection based on a joint assessment of semantic fit and salience. Ideally, a machine learning approach would automatically determine the weights such that the resolution accuracy is optimized.

The maximum entropy ranker that I will sketch here is closely related to similar approaches using probabilistic ranking for language modeling (Berger *et al.*, 1996; Rosenfeld, 1996) or parse selection (Johnson *et al.*, 1999), but uses the more efficient numerical optimization techniques that have become commonplace for maximum entropy learning in the meantime.

The most popular alternative to using a ranker for the selection of an antecedent in a *best-first* coreference resolver is to use the confidence of a classifier decision on positive/negative examples, with the benefit that almost all off-the-shelf classification methods can be used for such a purpose.

Many approaches in the literature (Morton, 2000; Hartrumpf, 2001; Ng and Cardie, 2002c; Luo *et al.*, 2004) use such a classification-based approximation to implement a best-first scheme.

Using the confidence of a positive/negative classifier as a choice, however, may not be the best way to realize antecedent choice; what we really want is a function that assigns a score to each possible antecedent (based on the features of the antecedent) *such that the correct antecedent frequently has the highest score*.

Seen from this perspective, the choice of using a binary classifier on single instances to learn this function is purely incidental, and more a result of the availability of ready-to-use machine learning packages (and the relative lack of easy-to-use packages that perform ranking) than a conscious choice of the researchers who used it.

To combine both salience indicators (sentence distance, and grammatical function) and indicators for similarity or compatibility (essentially those presented in the last chapter) into a score that can be used for ranking, each feature (which can be either binary or scalar) is multiplied with a corresponding weight.

In this fashion, each candidate can be represented as a vector of numerical features, and these feature values are multiplied with the feature weights to get the score of a candidate, and we can choose the candidate with the largest score:

$$\hat{y} = \arg\max_{y \in Y} \langle w, f(x, y) \rangle$$

(where w is the vector of the constraint weights, f is a function that maps a candidate to a feature vector, $\langle \cdot, \cdot \rangle$ is the dot product in Euclidean space, and Y is the set of possible antecedents).

To choose the constraint weights, we interpret our score in a probabilistic fashion. Given the measure[6]

$$\mu(x, y) := \mathrm{e}^{\langle w, f(x,y)\rangle}$$

we can define a probability distribution

$$\hat{P}(y|x) := \frac{\mu(x, y)}{\sum_{y' \in Y} \mu(x, y')} = \frac{\mathrm{e}^{\langle w, f(x,y)\rangle}}{\sum_{y' \in Y} \mathrm{e}^{\langle w, f(x,y')\rangle}}$$

by normalizing the measure so that the probabilities of all $y \in Y$ sum up to 1. This kind of model is called a loglinear model; in case of binary features, the feature weights can be interpreted as (the logarithm of) an odds ratio, whereas in the case of continuous features, the feature weights can be seen as the parameter of an exponential distribution.

Our loglinear model, given a weight vector w and a feature function f that computes features for a given alternative solution (antecedent candidate) $y \in Y$, will give us a means of choosing the most likely solution from Y. What is needed, then, is a way of finding good values for w by numerical optimization.

The most obvious way, choosing w such that the correct alternative comes out first in a maximally high number of examples, gives rise to an NP-hard optimization problem; thus, for machine learners such as AdaBoost, MaxEnt or Support Vector Machines, a *loss function* is chosen which (a) bounds the number of incorrectly classified examples from above, so that, all other things being equal, a lower loss value indicates a better weight vector and (b) is convex (which makes it possible to use efficient optimization procedures).

For maximum entropy learning, this loss function is the (negative logarithm of) the likelihood of the correct y according to the probability function $\hat{P}(y)$ we get from our loglinear model (frequently called the likelihood of the data given the model):

$$\begin{aligned} LL(w; \theta) &= \log \prod_{x,y \in \theta} \hat{P}(y|x) \\ &= \sum_{x,y \in \theta} \langle w, f(x, y)\rangle - \log \sum_{y'} e^{\langle w, f(x,y')\rangle} \end{aligned}$$

Together with a prior (an independently motivated probability distribution for w), the logarithm of the combined probability of model and data can be stated as a function $Loss(w)$:

[6] A measure fulfills all the requirements for a probability distribution, except that it does not need to be normalized. For finite sets, we can define a measure simply by assigning a non-negative number to every element (x,y) from the set $X \times Y$.

$$\begin{aligned} Loss(w) &= LL(w;\theta) + C \cdot ||w||^2 \\ &= \sum_{x,y\in\theta} \underbrace{\langle w, f(x,y)\rangle}_{\text{linear in } w} \underbrace{- \log \sum_{y'} e^{\langle w, f(x,y')\rangle}}_{\text{convex in } w} + \underbrace{C \cdot ||w||^2}_{\text{convex in } w} \end{aligned}$$

$Loss(w)$ is convex – given two weight vectors $w_1, w_2 \in \mathcal{R}^n$, any weight vector $w = u \cdot w_1 + (1-u) \cdot w_2$ on a line between w_1 and w_2 (i.e., with $u \in [0,1]$) will obey the equation

$$Loss(w) \leq u \cdot Loss(w_1) + (1-u) \cdot Loss(w_2)$$

The convexity of the $Loss$ function means that any local minimum – i.e., a value w^* such that $Loss(w^*) \leq Loss(w')$ for any w' in some open set U that contains w^* – is global: Suppose that w^* is not a global minimum and there exists a w'' such that $Loss(w'') < Loss(w^*)$. Then there exists an $u \in (0,1)$ such that $w' = u \cdot w^* + (1-u)w''$ is in the open set U. Because of convexity, $Loss(w') \leq u \cdot Loss(w^*) + (1-u) \cdot Loss(w'') < Loss(w^*)$, which would contradict the assumption that w^* is a local minimum of $Loss$.

The practical consequence of convexity is that any local optimization method can be used to find the minimum of the loss function.[7]

A large variety of optimization techniques can be used to numerically estimate the optimum weight vector for such a maximum entropy problem (see Malouf, 2002). I chose the most commonly used alternative, the L-BFGS algorithm (Liu and Nocedal, 1989), which uses a limited-memory approximation to the second derivative (the *Hessian*, which would be too expensive to store since its dimensionality is the square of the dimensionality of the feature space) to perform more effective updates than iterative scaling (the most commonly used algorithm in the 1990s), (non-stochastic) gradient descent, or conjugate gradient search.[8]

In coreference resolution, it is possible that we have multiple candidates that are all coreferent to the description we are looking at, and we do no longer have a single 'good' candidate y, but we can readily extend our model by simply considering the probability for *any one* of the correct candidates according to the model:

$$P_{\text{good}} = \sum_{y \in Y_{\text{good}}} \hat{P}(y|x) = \frac{\sum_{y \in Y_{\text{good}}} \mu(x,y)}{\sum_{y' \in Y} \mu(x,y')}$$

[7] This is to be seen in contrast with methods that allow for non-observed states in between such as multilayer neuronal networks and hidden-variable training using expectation maximization, where there is a significant probability that the training procedure converges to a local optimum that results in worse performance than the globally optimal parameter settings would yield, and the success of training such a model crucially depends on appropriate initialization and training regimes.

[8] The optimization code used is the L-BFGS routine by Liu and Nocedal (1989), which is available at `http://www.ece.northwestern.edu/~nocedal/lbfgs.html`. A wrapper generated by the F2PY program (Peterson *et al.*, 2001) allows the use of the optimization routines from Python.

The resulting loss function is no longer guaranteed to be convex, since it now contains a term $\log \sum_{y \in Y_{\text{good}}} \exp(\langle w, f(x,y) \rangle)$ which is concave. In principle, this could mean that there are different local minima w' and w'' where w' puts a higher weight on one $y' \in Y_{\text{good}}$ and w'' puts a higher weight on a $y'' \in Y_{\text{good}}$ which is different, with weight vectors in-between having worse values for the loss function than either of the two.

Such a situation may arise when there are multiple correct antecedent candidates in the window under consideration. For a hypothetical system that considers both string similarity and distance, local optima may arise that either assign a higher weight to string similarity or to distance. However, it is implausible that either of the local minima we get with the any-correct-candidate formulation would be worse than the single minimum that we would get when we limit ourselves to one positive candidate (for example, by only including the closest non-pronominal member of the coreference chain).

7.2 Resolving coreferent bridging

The heuristics for same-head resolution outlined in the last section, together with the evaluation results, show that it is possible to create a well-performing coreference resolver purely by considering same-head resolution.

However, a resolver for NP coreference should ideally be able to handle at least some cases that involve non-same-head coreference (which Vieira and Poesio call coreferent bridging), using an appropriate combination of information sources for semantic compatibility outlined in chapter 6 and for recognition of discourse-new noun phrases.

Of the 1556 anaphoric noun phrases in the training section of the corpus, only 1250 are resolvable with same-head techniques (of these 1250, about 70% are coreferring names, and the other 30% are definite descriptions). The rest can be split among several categories.

175 of the non-same-head cases are common nouns that can be resolved to another common noun (**NN→NN**), as in example (7.1) below:

(7.1) a. Lebensgefährliche Körperverletzungen hat sich [$_1$ eine 88jährige Fußgängerin] bei einem Zusammenstoß mit [$_2$ einem Pkw] zugezogen.
[$_1$ *An 88-year-old (female) pedestrian*] *has been gravely injured in a collision with* [$_2$ *a car*].

b. [$_1$ Die Frau] hatte [$_2$ das Auto] beim Überqueren der Waller Heerstraße offensichtlich übersehen.
When crossing the Waller Heerstraße, [$_1$ *the woman*] *had obviously overlooked* [$_2$ *the automobile*].

69 are common nouns that can be resolved to a name (**NN→NE**), as in example (1.6), repeated here as (7.2):

(7.2) a. "Wir müssen uns selbst helfen", meinte [$_3$ Magath].
"We have to help ourselves", said [$_3$ *Magath*].

b. [$_3$ Der Werder-Coach] bangt um die Einsätze von Detlef Eilts und Marco Bode.
[$_3$ *The Werder Bremen coach*] *is worried about Detlef Eilts and Marco Bode's ability to play*.

25 are names that must be resolved to a different name (**alias**), such as *Volkswagen–VW* in example (7.3):

(7.3) a. Die drei rockmusizierenden älteren Herrschaften von [$_2$ Genesis] verbanden sich mit den autofabrizierenden älteren Herren von [$_1$ Volkswagen], um fürderhin für gegenseitige Belebung des Geschäfts zu sorgen.
The three rock-music-making elderly gentlemen from [$_2$ *Genesis*] *joined the car-manufacturing elderly gentlemen from* [$_1$ *Volkswagen*] *to henceforth take care of the mutual stimulation of their businesses.*

b. 20 Millionen läßt sich [$_1$ VW] das Sponsoring von [$_2$ Genesis] kosten, und die haben dafür gar keine Probleme damit, daß neuerdings [$_3$ ein Auto ihres Namens] ökologische Probleme verschärft.
For the sponsoring of [$_2$ *Genesis*], [$_1$ *VW*] *lays out 20 million (DM), and in turn, these have no problems with the fact that* [$_3$ *an automobile named after them*] *recently set out to aggravate ecological problems.*

c. [$_3$ Das VW-Modell Genesis] ist dementsprechend ähnlich aufregend wie die inspirierende Musik.
Accordingly, [$_3$ *the VW model Genesis*] *is just as exciting as the inspiring music.*

Three mentions are names that must be resolved to a noun, such as this example of a **cataphoric specific indefinite** in (7.4):

(7.4) a. [$_1$ Spanischer Ölkonzern] surft auf Fusionswelle.
[$_1$ *Spanish oil company*] *surfs on wave of mergers.*

b. [$_1$ Repsol] will hoch hinaus.
[$_1$ *Repsol*] *is aiming high.*

In such cases, an entity is first introduced by a common noun phrase, only to be mentioned again using the full name.

26 mentions have a potential same-head antecedent but have to be resolved indirectly, as in the mention "*das VW-Modell Genesis*" (*the VW model Genesis*) in example (7.3), which shares a name component with *Genesis* (the band), but in fact has to be resolved to the indefinite "*ein Auto ihres Namens*" (*a car named after them*).

Eight mentions have an indirect antecedent that disagrees in number (this is the case for e.g. a team that is metonymously mentioned as "the players")

(7.5) a. Nach beiden Ausgleichstreffen [$_1$ der Berliner] wurden auf der Anzeigetafel Führungstore der Mönchengladbacher in München vermeldet.
After the two equalizers of [$_1$ *the Berliners*], *the lead goals of the Mönchengladbach team were registered on the scoreboard in Munich.*

b. [$_1$ Berlin] spielte nach einem Platzverweis sogar noch eine Viertelstunde in Unterzahl.
[$_1$ *Berlin*] *even played short-handed for a quarter of an hour after receiving a sending-off.*

Seen with respect to all definite descriptions and names, or even all discourse-old full noun phrases, the coreferent bridging cases are in the minority. They consitute slightly less than half of definite nominals, or about 20% of all discourse-old full NPs including names. However, trying to resolve them can both yield linguistically interesting insights and it may also be possible to improve both precision and recall – in comparison to doing only same-head resolution – if we are able to resolve coreferent bridging cases with rather high precision. In the rest of this chapter, I will present some experiments where the same-head resolver is complemented with a component based on the information sources presented in chapter 6 that may help in cases like example (7.1).

The referents 1 (the woman) and 2 (the car) in example (7.1) are mentioned again in the second sentence, not pronominalised, nor repeated identically, but in a semantically poorer form (the [female] pedestrian - the woman), or as a synonym (the car - the automobile). In contrast to pronominal reference or same-head coreference, it is possible that anaphor and antecedents have a differing grammatical gender ('Pkw' has male gender, while 'Auto' has neuter gender), whereas number disagreement is much less frequent (it is possible in cases like (7.5) where an organization is metonymously referred to by a plural person reference).

In the case of non-same-head resolution, the choice of an appropriate antecedent is much more difficult (in chapter 6, we see that, given a discourse-old definite description, the closest same-head antecedent has a 87% chance of being the correct antecedent, whereas GermaNet hyperonymy can pick out the correct antecedent only 67% of the time, and lower-precision methods only being right about 35-45% of the time), while the large majority of all definite descriptions without a same-head antecedent is simply discourse-new and should never be resolved.

Indeed, many mentions which we would be able to resolve in the setting of chapter 6, given the information that they are discourse-new, are out of reach in the more realistic setting. Consider the following example (7.6):

(7.6) a. [$_3$ Nikolaus W. Schües] bleibt Präsident der Hamburger Handelskammer.
[$_3$ *Nikolaus W. Schües*] *remains president of the Hamburg Chamber of Commerce.*

b. [$_3$ Der Geschäftsführer der Reederei "F. Laeisz"] wurde gestern für drei Jahre wiedergewählt.
[$_3$ *The managing director of the shipping company "F. Laeisz"*] *was reelected for three years yesterday.*

In this example, "*Geschäftsführer der Reederei 'F. Laeisz'*" (managing director of the shipping company "F. Laeisz") is uniquely specifying: if the noun phrase oc-

	P_{max}	R_{max}	P_{min}	R_{min}	Perp	Prec	Recl	F
features for coreferent bridging								
no filter	62.3	92.5	14.3	61.6	1.42	62.0	68.6	65.1
+gwn only	62.3	92.5	14.3	61.6	1.28	62.0	68.6	65.1
unique_mod	60.7	86.3	21.2	61.6	1.51	62.0	68.6	65.1
+segment	60.6	85.6	21.4	61.6	1.49	62.0	68.6	65.1
+num	60.6	85.6	21.4	61.6	1.49	62.0	68.6	65.1
+gwn	59.8	83.0	21.7	61.6	1.28	61.7	69.2	65.2
+syn_role	59.8	83.0	21.7	61.6	1.27	61.9	69.5	65.5
NE_semdist	59.8	83.0	21.7	61.6	1.27	61.9	69.7	65.6
+pred_arg	59.8	83.0	21.7	61.6	1.26	61.9	**70.0**	**65.7**
comparing learning methods								
all-candidates (same as +pred_arg)						61.9	70.0	65.7
unary learning scheme (binary yes/no classifier)						61.0	69.1	64.8
binary learning scheme (choice between two cand.)						61.8	69.2	65.3

gwn: node distance in GermaNet
syn_role: syntactic role of the potential anaphor
NE_semdist: node distance in GermaNet for NEs
(replace NE antecedent with node for semantic class)
pred_arg: use model of selectional preferences

Actual precision (Prec,Recl) has been determined on the evaluation dataset; upper/lower bounds on precision and recall (Pmax,Pmin,Rmax,Rmin) were determined on the training dataset.

Table 7.2: Upper and lower bounds / coreferent bridging

curred as a discourse-new mention, it would still receive a definite article. Therefore, it would not make sense to resolve the definite description in (7.6-b) to an antecedent unless we were able to use the world knowledge that Nikolaus Schües is in fact the managing director of that shipping company. A human reader would be able to use knowledge about discourse structure to force the connection (the fact that someone has been reelected in a position is somewhat irrelevant unless we know who it is, and while '*managing director of F. Laeisz*' is uniquely specifying, it is not informative enough for the reader of the newspaper).

Therefore, the most promising strategy to resolve coreferent bridging in a realistic setting would be to focus on a subset of the discourse-old cases – essentially those of the NN→NN and NN→NE cases where we have a chance of finding a correct antecedent with high probability. A majority of these cases have good locality properties: of the 175 NN→NN cases, 141 are within four sentences, and of the 69 NN→NE cases, 68 are within four sentences of the definite description.

Targeted subset The approach chosen in TUECOREF aims at resolving a well-behaved subset of the coreferent bridging cases by using a high-precision approach to find likely antecedents for those definite descriptions that cannot be resolved in the same-head resolver.[9]

The first step involves a hard filtering step to eliminate those mentions that most likely should not be resolved:

- All names (i.e., mentions that contain a head word which is tagged as NE) are treated as unique.
- All kinds of postmodification (genitive postmodifiers, prepositional phrases, relative clauses as well as postmodifying infinitive clauses) are seen as uniqueness indicators.
- Noun phrases with ordinal adjectives (first, second, etc.) or invariant adjectives (*Berliner*, *Hamburger*, ...) are also treated as discourse-new.

For any definite description that passes this test, all full noun phrases are selected as potential antecedents if they are within the last four sentences (antecedents within the current sentence are disregarded) and filtered based on number agreement. If the definite description to be resolved has *person* as its semantic class, additional filtering is performed based on gender agreement.

Among these potential non-same-head antecedents and a pseudo-candidate that indicates that the mention is discourse-new and should not be resolved, a maximum entropy ranker chooses the most likely alternative based on informative features.

With only GermaNet distance, or only the semantic classes of anaphor and antecedent, the system is not confident enough about a possible antecedent (see table 7.2, rows labeled *no filter*, *unique_mod* up to *+num*). However, the perplexity shows a marked decrease using the information from GermaNet, indicating that the GermaNet distance is indeed an informative feature, even though no additional coreference links are found.

A set of features including **GermaNet distance** and the pair of **semantic classes** of the definite description and potential antecedent (see the row labeled *+gwn* in the *unique mod* group of table 7.2) results in a visible improvement in recall, at the cost of a much smaller drop in precision.

In order to improve the discourse-old/discourse-new information in the resolver, I also include the **grammatical role** of the definite description, as it can provide an indication towards its information status (for example, subjects tend to be discourse-old). Besides verb arguments (subjects and accusative/dative objects), as well as genitive postmodifiers, the approach distinguishes between one group of prepositions (*aus, bei, für, gegen, seit, um*) which frequently governs discourse-old

[9] For the non-same-head experiments, I use the larger-recall *unique names* variant rather than the variant with loose segmentation, as the recall lost through the segmentation heuristic would be difficult to substitute using non-same-head links.

noun phrases, and other prepositions, which are more likely to govern a discourse-new noun phrase. The inclusion of this feature (given as *+syn_role*) gives a small boost in terms of both precision and recall for the system, together with a small improvement in the perplexity value.

As a further step, I use a modified GermaNet distance feature that differentiates **hyperonymy** proper and semantic similarity, i.e. the semantic distance feature is split into one for hypernyms (where the anaphor is synonymous or more general than the antecedent), one for general semantic distance (in the case that there is no hypernymy relation between anaphor and antecedent).

Additionally, named entities that are not found in GermaNet are represented by a general term corresponding to the semantic class (e.g. the *person* synset for person NEs), further increasing recall by a small amount. (see the *NE semdist* row in table 7.2).

The result including these features is a visible improvement in recall – from 68.6% to 69.7%, or a 1.1% absolute improvement – at a rather small cost in precision (from 62.0% precision to 61.9%).

Comparing learning methods The *all-candidates* scheme, where a ranker is trained to select the best candidate among the complete list of antecedent candidates, was used in the work described here, and earlier in (Versley, 2006), as well as in the experiments of Denis and Baldridge (2007b) using a maximum entropy ranker (for non-same-head definite descriptions, and pronoun resolution, respectively).[10] Denis and Baldridge directly compare unary, binary and all-candidates approaches and find that the all-candidates approach is clearly superior.

In the context of non-same-head resolver, the following schemes for machine-learning-based selection of an antecedent can be used:

- The *unary* scheme is the one mentioned above: for all positive instances, we want a positive weight, and for all negative instances (i.e., incorrect antecedents), we want a negative weight.
- In a *binary* scheme, one would require that each positive instance gets a greater weight than any negative instance for the same anaphoric description.[11]
- In an *all-candidates* scheme, one would want that at least one of the positive instances has a weight that is greater than any negative instance for the same anaphor.

[10]The online learning approach of Daumé III and Marcu (2005) is another example of the all-candidates approach.

[11]The binary scheme most closely corresponds to the approach of Yang *et al.* (2003). However, Yang *et al.* incorporate surface ordering by always considering pairs of antecedents, which means that their approach can learn a *closest antecedent with property X*-style resolution rule, but cannot easily model confidence.

non-resolution		*resolution*	
non-resolution bias	0.26	PER→PER	1.61
new: PER,sg	–1.83	ORG→ORG	1.63
new: ORG,sg	0.16	LOC→LOC	0.92
new: LOC,sg	–0.45	EVT→EVT	0.84
new: TMP,sg	0.72	TMP→TMP	–0.18
new: SUBJ	–0.97	sentence distance	–0.75
new: PP	0.95	GWN node distance	–0.60
		GWN dist (NE)	–0.65
		Pred-Arg odds	–0.83

Table 7.3: Some of the constraint weights learned for coreferent bridging

If I compare the results of choosing the weights using the ranker described earlier (all-candidate approach) with the binary-classifier approximation mentioned earlier (using incorrect antecedents as negative examples, and correct ones as positives), I find that the ranker performs significantly better. Using exactly the same features as the best-performing system (using GermaNet, grammatical roles, and the selectional preferences model) with the *unary* classification scheme would result in worse precision (61.0%, vs. 61.9% using a ranking classifier), but also visibly lower recall (69.1% vs. 70.0% in my system).

Constraint weights One advantage of the maximum entropy ranker over other competitive learning models is that its feature weights are interpretable and can be used to uncover the relative importance of the information sources.[12]

In the model used for coreferent bridging (see table 7.3), we see that both sentence distance and semantic distance are important factors. There are quite large differences in the behaviors of the different semantic classes with respect to (non-)anaphoricity, with persons having a large preference to being anaphoric (both in terms of resolution, where resolving a person anaphor to a person antecedent carries a large positive weight, and in terms of introducing a new referent, where introducing a new person referent carries a large negative weight). We also see confirmed Prince's (1992) observation that subjects tend to be discourse-old.

7.2.1 Beyond semantic similarity

Generally speaking, the model still suffers from a lack of information, be it from the inaccuracy of semantic classes, lack of correspondence to GermaNet's topology in the case of quasi-synonyms, word sense ambiguity.

[12]Statistical significance tests, which are straightforward to do in simple – low-dimensional, unregularized – logistic regression models, would require taking into account both the regularization and correcting for multiple testing in addition to computing leave-one-out or cross-validation estimates for likelihood. As this would go beyond the scope of this thesis, I will limit myself to an exploratory discussion of the most important weights as learned by the classifier.

The surrounding context can provide additional information that is not apparent from the mention heads alone. In chapter 6, information from the context was not considered at all (mostly because only one source of information was considered at the same time, and no attempt was made to combine them beyond the most-precise-first approach). Using the MaxEnt ranker for combining complementary information sources, however, it is possible to take into account such context information.

A model for selectional preferences For my system, I decided to use a rather minimalist approach, trying to capture some of the generalizations made by Bean and Riloff's model but simpler and presumably more robust in nature. For noun phrases which are either the subject or the accusative object of a verb, I use a statistical model of selectional preferences to assess whether a potential antecedent and the context of the anaphoric NP are compatible.

Using the automatic parses from "die tageszeitung" which also served to create the distributional similarity measures mentioned in chapter 6, I extracted subject-verb and direct object-verb pairs and I trained models for both relations using LSC, a soft clustering software by Helmut Schmid[13] based on the Expectation Maximization algorithm. Compounds were split with the unsupervised algorithm described in section 6.2.3 that takes into account both common and productive prefixes and the general frequency of the word, so that more frequent compounds are left intact. This is an advantage in cases where the base noun has several senses, but a compound is unambiguous (for example, the German word *Organization* describes both an organizational body and a process of organising something, whereas the compound *Nachfolgeorganization* is only used in the organizational body sense).

A feature based on selectional preferences should serve to detect two cases: One is that an antecedent is implausible, in which case the probability (according to the model) to see the definite description in the antecedent's context would be lower, and the model probability of the antecedent in the definite description's context would also be lower. The other case would be that either the definite description or its antecedent are part of a non-compositional collocation (such as *take place*, *take aim*).

Intuitively, the (pointwise) mutual information between the potential antecedent and the context of the mention to be resolved should be as large as the mutual information between the antecedent and its original context, or the mention to be resolved and its context, and similarly for the (pointwise) mutual information between the potential antecedent's context and the mention to be resolved.

In statistical terms, the mutual information is defined as (the logarithm of) the ratio between the (actual) joint probability and a probability computed under independence assumption. The difference between two mutual information values

[13] http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/LSC.html

corresponds to (the logarithm of) the ratio between the two corresponding joint probabilities.

This gives us a useful statistic that (i) reflects the (semantic) plausibility of exchanging a definite description with a given candidate antecedent and (ii) gives a continuous value that reflects how certain we can be that exchangeability is not the case: in the case where either the antecedent does not fit in the context of the definite description or the definite description does not fit in the context of the antecedent, the statistic takes a more negative negative value as the likelihood that for the switched pair to occur would be much smaller.

The statistic can be expressed as follows:

$$q := \log \frac{p_{r'}(n_1, v_2) \cdot p_r(n_2, v_1)}{p_r(n_1, v_1) \cdot p_{r'}(n_2, v_2)}$$

If the mention to be resolved is very likely to appear in the antecedent's context and vice versa, q should be near (or even above) zero, while a negative value of q indicates that the anaphor does not fit to the antecedent's context or vice versa. Positive values are clipped at +0.0, as there is no obviously sensible interpretation other than that anaphor and antecedent are exchangeable (which is, however, already the case at a value of 0.0).

As an example, consider *Arbeiterwohlfahrt* (a German charity organization) as subject of *entlassen* (to lay off) and *Mark* (currency) as subject of *fliessen* (to flow), which are not exchangeable and yield a large negative value (-5.9) since the switched version is about 370 times less likely than the original one. In contrast, *Siegerin* (victor) as object of *disqualifizieren* (disqualify) and a person name as subject of *landen* (to land) are well exchangeable, yielding a positive value (+1.0).

In table 7.2, the addition of this feature corresponds to the row labeled *+pred_arg*. While precision stays at the same level, adding the feature results in slightly improved recall (from 69.7% to 70.0%).

Corpus-derived and web-based features In the discussion in chapter 6, the most successful combination of indicators for non-same-head antecedents include GermaNet hyperonymy, which is used here together with the GermaNet distance feature, but also the use of web-based pattern search, the 25-most-similar items according to the distributional similarity measure, and a technique combining distributional similarity with distance (rejecting an antecedent candidate when a mention further away has a greater distributional similarity value).

Of these information sources, only the most precise one (web patterns) had a positive effect, whereas the use of the 25-most-similar items list did not improve the results. Table 7.4 summarizes results using these features.[14]

[14] Due to the fact that a different release of the TüBa-D/Z data was used, precision and recall figures differ from the other tables by 0.3% for precision and 0.2% for recall, resulting in an F-measure that is 0.1% lower than those in tables 7.1 and 7.2.

	Prec	Recl	F
unique_mod (same-head)	61.7	68.8	65.1
gwn+semclass	61.6	69.8	65.4
+pred_arg	61.6	70.1	65.6
+web patterns	62.0	**70.7**	**66.1**
+n-most-similar (cf. Gasperin and Vieira, 2004)	**62.1**	70.6	**66.1**
recall-oriented	59.3	**72.4**	65.2
precision-oriented	**80.2**	54.5	64.9

Table 7.4: Addition of corpus-derived/web-based features

The information from web patterns yields an absolute improvement of 0.6% recall and 0.4% precision; the distributional thesaurus approach (see the row labeled n-most-similar in table 7.4) yields a slight improvement in precision accompanied by a slight deterioration of recall.

Different Precision-Recall Balances Varying with applications, and more so for the resolution of definite descriptions where results have certain imperfections, it may be useful to tune the resolver to either higher recall (e.g., to get more suggestions in an automatic annotation tool) or higher precision (e.g., to prevent overmerging in an information extraction task).

As noted earlier, the precision for name matching is relatively high, and the precision for the coreferent bridging component is slightly above that for same-head resolution for nominals. To improve the recall, it is possible to artificially lower the threshold at which non-same-head coreference is resolved by decreasing the weight of the non-resolution item in the ranking. Adding a fixed term of -2 to the non-resolution weight, recall raises from originally 70.1% to 72.4%, while precision goes down from 61.6% to 59.3% – a rather balanced tradeoff which incurs only a minimal decrease in F-measure (from 65.6% to 65.2%).

To improve the precision, it is possible to introduce a model for same-head resolution that takes into account the distance distribution for different semantic classes, yielding a ranking model that is similar to the non-same-head ranker, but relatively simpler (since the head equality obviates the need for semantic similarity measures). Using this model, it is possible to improve the precision to 80.2%, while recall decreases to 54.5%.

7.2.2 Discussion of the Evaluation Results

The version of my system that includes resolution of coreferent bridging by means of semantic classes, GermaNet distance, selectional preferences, and salience indicators such as sentence distance and grammatical function, has 70.0% recall. This is 1.4% above that of same-head resolution, whereas precision is lower by only 0.1%. To put this in absolute terms, 32 more definite descriptions (of 1340) have

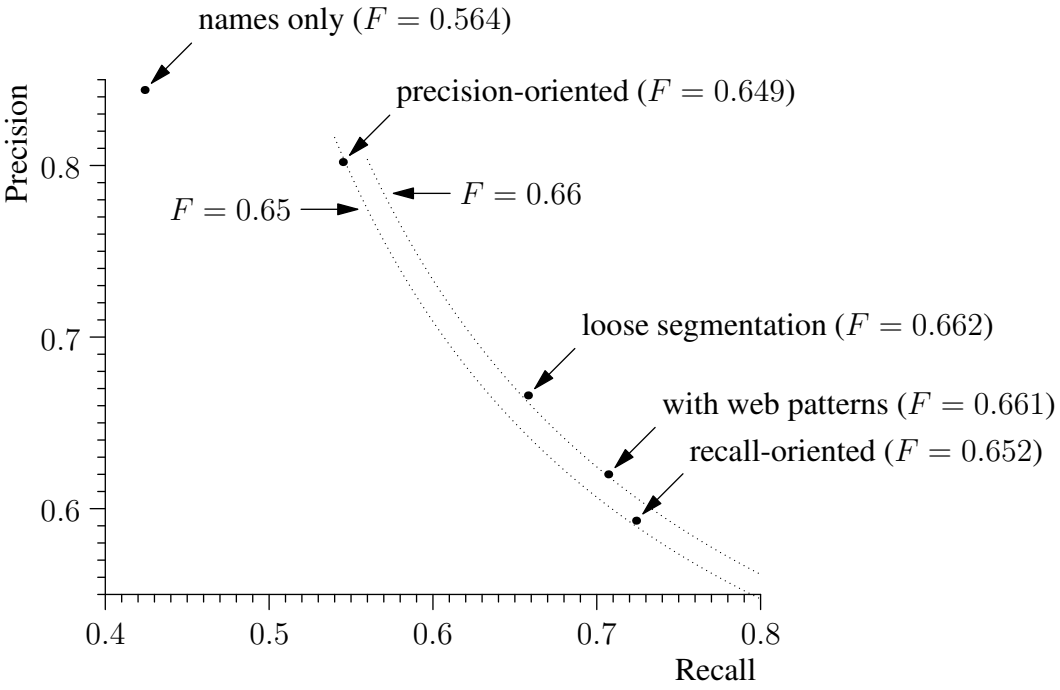


Figure 7.1: Precision-Recall balance of different results

been resolved and the correct antecedent was found for 17 of them, giving a precision of 53% for coreferent bridging, which is quite close to the same-head precision for common nouns (albeit with a much lower recall).

The version that uses web-based pattern search in addition shows a further improvement in recall, to 70.7%, or about 48 more definite descriptions than the same-head resolver; trading in some precision for recall, it is possible to reach 72.4% recall at a precision that is still above the precision of the "head identity" baseline in table 7.1. Thus, the model is relatively successful in improving recall while keeping a high precision.

Compared to the same-head systems, we can see that there is clearly a tradeoff to be made between higher precision and higher recall (the *loose segmentation* version of the same-head resolver reaches an F-measure very close to the coreferent bridging variant including web patterns), but that the added complexity of coreferent bridging resolution results in substantially decreased resolution speed, even using high-performance implementations for semantic similarity and wordnet search. In sum, the research presented in this chapter shows that resolution of coreferent bridging can be done with enough precision to be useful in a real-world setting, but the desirability of doing so still depends on an application perspective. Both system variants – the same-head version as much as the version that achieves better recall by resolving coreferent bridging cases – achieve a good tradeoff between precision and recall. More importantly, the improved same-head resolution

(which achieves 84.6% precision and 90.6% recall in the resolution of names) is useful independently of the approach used for definite descriptions.

Finally, we can compare the results for the complete-system case – if we consider only same-head resolution using the *unique names* approach, we get 49.4% recall and 43.8% precision for the realistic setting, against 87% precision for the antecedent selection task – rougly half the precision. As a result, relatively noisy approaches such as ranking by distributional similarity appear highly attractive in the antecedent selection task, whereas they would invariably hurt the performance in the more realistic task. In the realistic task, distributional similarity measures fail to be beneficial to system performance despite measures to improve precision such as filtering by distance (remember that we only resolve to antecedent candidates that are at most four sentences back) and filtering out some spurious matches using the selectional preferences model.

These results point to the importance of an effective integration of discourse-new indicators with resolution itself, a problem that remains to be solved despite encouraging progress by current state-of-the-art approaches to coreference resolution. Similar to the situation with antecedent selection, we see here a disconnect between relatively high results in oracle settings (Uryupina, 2003 achieves an F-measure of 92.0% for uniqueness classification) yet the results of actually using such classifiers make them sound much less appealing, due to the fact that they significantly decrease recall. Uryupina (2007, p. 185) summarizes this as "*... error-prone automatically induced detectors are not reliable enough to produce a similar [to the ideal detector] precision gain and the system's F-score goes down because of the recall loss, as the baseline's recall is already relatively low.*"

The solution chosen in the experiments of this chapter – use of the maximum entropy ranker with a pseudo-candidate that allows the ranker to mark a case as discourse-new – is successful insofar as it achieves an acceptable precision for non-same-head resolution of *easy* cases (those where there is a near-synonymy, hyperonymy or instance relation that can be detected) but is not sufficient to yield a recall anywhere close to the more optimistic results in chapter 6.

Chapter 8

Conclusions

I have presented the TUECOREF system for resolution of names and definite descriptions in German newspaper text, and the components it uses. TUECOREF uses a large variety of information sources – GermaNet, pattern search on the World Wide Web, grammatical function and sentence distance of a potential antecedent, distributional similarity and association measures – to make optimal use of the information inherent in the context of the definite description or name to be resolved.

Features are complementary: they cover different aspects of the context that are relevant, such as relations between the heads of mentions, selectional preferences, and salience as expressed through sentence distance and grammatical function; or they pertain to different types of antecedents (common noun antecedents can be detected using the lexical information contained in GermaNet, whereas name antecedents are covered better by pattern-based search in the World Wide Web). Different features also form a continuum that ranges from precise methods with lower recall to methods that offer larger improvements in recall at the cost of lower precision.

The experiments, which use data from the referentially and syntactically annotated TüBa-D/Z corpus, aim towards answering two distinct but related questions: One is *whether we have sufficient information for finding an antecedent to a discourse-old definite description*. The other question regards *how to construct a system that performs resolution of definite descriptions and names accurately*.

The first question can be given a mostly positive answer: It is possible to resolve definite descriptions with 68% recall overall; the precision that is reached over all definite descriptions (same-head or not) is 79%. Looking at same-head and non-same head descriptions separately, the system achieves 87% precision for same-head antecedents and 63% precision for coreferent bridging (i.e., non-same-head) antecedents.

Regarding the second question, it is important to note that the accuracy of a system hinges on accurate same-head resolution of names and definite descriptions, since these cases are by far the majority of the resolvable non-pronouns. Using a rule-based approach that incorporates heuristics for name structure (and

its variation) as well as variation of modifiers in same-head definite descriptions, TUECOREF reaches 84.6% precision and 90.6% recall for names, and 43.8% precision at 49.4% recall for definite descriptions, or 62.2% precision and 68.6% recall overall.[1]

Additional improvements on the same-head version of the system are reached by resolving a subset of non-same-head definite descriptions, integrating features for discourse-new detection with informative features for the choice of an antecedent.

The usefulness of high-recall, low-precision information sources such as those based on unsupervised learning from corpora depends on the setting that is assumed: In a more forgiving setting where the definite description to be resolved is always discourse-old, an approach using a generic distributional similarity measure yields a substantial improvement in recall. In a setting where the system has to solve the dual task of deciding whether a given definite description is discourse-old or discourse-new *and* choosing an antecedent, low-precision features do not yield a visible advantage since it is still more likely that the definite description under consideration is discourse-new than it is that the description has the predicted antecedent.

Developing a coreference resolution system for German poses challenges in several specific areas that are not present in English coreference resolution: A very basic one is German morphology, which means that head matching will need to be complemented by morphological analysis and/or lemmatization to be maximally useful. A more pervasive problem for German is its synthetic compounding: head matching techniques, but also the use of semantic resources such as GermaNet or the creation of distributional semantic models make it necessary to take into account compounds. In the case of GermaNet, taking into account compounds is necessary to achieve optimal coverage. The use of distributional similarity models benefits from a frequency-sensitive compound splitting technique to avoid sparse data effects. As a final point, the resource situation for German is different from that for English. Due to different populations of speakers of the respective language, the German-language portion of the World Wide Web is considerably smaller than the English-language portion; furthermore, techniques for parsing or other methods of extracting grammatical relations are generally slower or less accurate for German than for English, which is why additional attention to the respective components is necessary in comparison to English.

In the existing literature, the problem of finding coreferent bridging antecedents to definite descriptions through a well-defined selection of informative features has been treated by Vieira and Poesio (1997), Markert and Nissim (2005) and Garera and Yarowsky (2006), using WordNet relations, web pattern search, and unsupervised learning from corpora, respectively. Additionally, work such as Poesio *et al.*

[1] Note the decrease from 87% precision to about 44% precision due to discourse-new definites. Current approaches to filter out these discourse-new definites with a potential same-head antecedent are largely unattractive because they result in a considerable recall loss.

(1998), Poesio *et al.* (2004a) or Goecke *et al.* (2008) aims at resolving bridging references in general without making a distinction between coreferent and non-coreferent antecedents.

In comparison to such work, the present thesis makes the following improvements: Firstly, it increases the recall of Web pattern search by combining the information from multiple patterns in an informative way. Such a combination approach using smoothed mutual information estimates outperforms every single pattern in terms of precision and recall since it is more tolerant of spurious matches, but also profits from single-pattern matches for low-frequency items.

In the area of distributional similarity and association metrics, the present work is the first to compare different similarity and association metrics with regards to the antecedent selection task. In the setting that was investigated here, a generic distributional similarity approach offers better accuracy than the specialized method that has been proposed by Garera and Yarowsky (2006). Existing work using association measures is either not concerned with distinguishing coreference from non-coreference (Poesio *et al.* or Goecke *et al.*) or does not consider the possibility of non-coreferent antecedents at all (Garera and Yarowsky). The method proposed in section 6.4 combines evidence from a distributional measure with recency and a semantic class filter to achieve much improved precision at a relatively small cost in terms of recall.

With respect to work on systems for the resolution of definite descriptions and names for German, both Hartrumpf (2001) and Strube *et al.* (2002) tackle the task of resolving non-pronoun mentions. Hartrumpf only gives evaluation figures for all markables (including pronouns and non-pronouns); the licensing rules that Hartrumpf uses are only partially listed, and the system uses the proprietary semantic lexicon HagenLex, which means that the approach cannot be easily reimplemented. Strube *et al.* (2002) use a system based on a decision tree classifier and report separate figures for personal and demonstrative pronouns as well as for common noun phrases and names. While they test their approach on a different text genre, it is described sufficiently well that it allows a reimplementation of the approach. As described in subsection 7.1.3, the reimplementation of Strube *et al.*'s decision-tree system performs on roughly the same level as the results reported in their paper. In the default setting, TUECOREF offers substantially more recall than the decision tree baseline. In a configuration that trades some recall for precision in a more informed way (the *precision-oriented* variant discussed at the end of section 7.2.1), it offers substantial improvements in both recall and precision over the decision tree baseline.

A number of issues present themselves for future research. On the side of information sources, this thesis has focused on unsupervised learning, as supervised learning of the semantic relations between an anaphoric definite description and its antecedent without any generalization would run into sparse data problems. Methods for semi-supervised learning, which are able to use a combination of labeled and unlabeled data, may be able to achieve a coverage close to that of the

unsupervised approaches while being better adapted to the task at hand. Both the referential annotation itself and existing resources such as GermaNet would be conceivable sources for labeled data.

One example for a fruitful use of both labeled and unlabeled training data would be a method that combines the intuition behind Garera and Yarowsky's (2006) unsupervised approach with a supervised classifier to yield better precision. Garera and Yarowsky use a mutual information statistic in conjunction with possible non-same-head antecedents of an anaphoric definite description. Weighting these pairs to reflect additional information (for example, using the relative ordering assigned by the ranking resolver) would result in a cleaner input distribution for the association statistic.[2]

With respect to the source of unannotated text, the research reported here uses the TüPP-D/Z corpus, which consists of 200 million words from the German newspaper `die tageszeitung`. Using text from the same domain (i.e., the same newspaper) is obviously preferable to using out-of-domain or lower-quality text. However, using a larger amount of text (which would be possible with DE-WaC, a collection of texts collected from the World Wide Web) may still be preferable despite the out-of-domain nature of the texts since using more data may yield better results in the unsupervised learning approaches. In sum, this is a question that can only be answered empirically. Among the unsupervised approaches examined in this thesis, the relation-free distributional similarity measure of Padó and Lapata (2007) only requires cheap unlabeled dependency parsing and could be applied to very large text collections (i.e., in the order of billions of words) without incurring disproportionate computational requirements.

The approach using pattern search on the World Wide Web offers good coverage on instance relations, which is instrumental in the treatment of definite descriptions with named entity antecedents. At the same time, it is the most expensive of the approaches proposed here, especially with the essential ingredient of combining multiple patterns. For a coreference system where throughput is a concern, it would therefore be preferable to have a cheap replacement with similarly advantageous qualities regarding precision and coverage. Google's N-gram dataset, which recently has been released in a version for German[3] may be a viable alternative to web queries if it were possible to counterbalance the smaller size (the whole German World Wide Web versus a 100 billion words sample) by using more patterns. Another, very different, approach would be to collect specialized gazetteer lists for definite descriptions which frequently have named antecedents (e.g., *the author*, *the city*, *the country*, *the party*, *the company*, *the newspaper*), possibly automating the process of searching for appropriate input to the gazetteer lists.

[2] Note that the co-training approach of Müller *et al.* (2002) only uses features that can be learned using relatively few training examples, which means that only a limited improvement would be possible. In contrast, learning a weighting for pairs of head lemmas is not possible using only annotated data, while a reasonably accurate classification for such pairs is highly desirable.

[3] Web 1T 5-gram, 10 European Languages, Version 1; LDC2009T25

Finally, the experiments reported in this thesis make use of the gold-standard syntactic and morphological information that is in the TüBa-D/Z treebank.

An approach that does coreference resolution from raw text (which would be the more common case for the actual application of a coreference resolver in a larger system) may suffer from some of the imperfections in the preprocessing components. The only morphological distinctions that are crucial for the resolution of definite descriptions are number (generally) and gender (for persons). Despite non-perfect morphological tagging, these distinctions can be reproduced rather accurately. In contrast, failures to distinguish common nouns and names in part-of-speech tagging (NN and NE part-of-speech tags) happen more frequently in part-of-speech tagging. These errors could be more problematic as common noun phrases and names are treated differently in resolution. Finally, the attachment of postnominal modifiers (prepositional phrases and relative clauses), which is important for the identification of discourse-new mentions, frequently shows errors in state-of-the-art parsers.

In the SemEval 2010 task on coreference resolution (Recasens *et al.*, 2010), participants' results showed a marked performance decrease between gold-standard preprocessing and system-created mentions (between 10% and 20% for many systems on both CEAF/ECM and MUC scores). Improving preprocessing components by targeting especially those parts of morphological tagging and parsing that are critical to mention identification and coreference resolution may be instrumental in improving the performance of coreference resolution system when using system-created mentions.

Bibliography

Abbott, B. (1999). Support for a unique theory of definite descriptions. In *SALT 9*.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, **7**(1), 131–177.

Almor, A. (1999). Noun phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, **106**(4), 748–765.

Almuhareb, A. and Poesio, M. (2005a). Concept learning and categorization from the Web. In *Proc. of Annual Meeting of the Cognitive Science Society (CogSci 2005)*.

Almuhareb, A. and Poesio, M. (2005b). Finding concept attributes in the Web. In *Proc. of the Corpus Linguistics Conference*.

Aone, C. and Bennett, S. (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proc. ACL 1995*.

Ariel, M. (1990). *Accessing Noun Phrase antecedents*. Routledge, London.

Ariel, M. (2001). Accessibility theory: An overview. In *Text representation*. John Benjamins, Amsterdam.

Ariel, M. (2007). A grammar in every register? The case of definite descriptions. In *The Grammar-Pragmatics Interface: Essays in Honor of Jeanette K. Gundel*, pages 265–292. John Benjamins.

Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

Asher, N. (2006). Things and their aspects. *Philosophical Issues*, **16**(1), 1–23.

Asher, N. and Lascarides, A. (1998). Bridging. *Journal of Semantics*, **15**(1), 83–113.

Asher, N. and Pustejovsky, J. (2005). Word meaning and commonsense metaphysics. http://semanticsarchive.net/Archive/TgxMDNkM/.

Attardi, G. and Ciaramita, M. (2007). Tree revision learning for dependency parsing. In *Proc. HLT-NAACL 2007*.

Bagga, A. and Baldwin, B. (1998a). Algorithms for scoring coreference chains. In *Proceedings of the First International Conference on Language Resources and Evaluation (LREC 1998)*.

Bagga, A. and Baldwin, B. (1998b). Entity-based cross-document coreferencing using the vector space model. In *Proc. ACL 1998*.

Bagga, A. and Bierman, A. (1997). Analyzing the complexity of a domain with respect to an information extraction task. In *Proceedings of the 10th Research on Computational Linguistics International Conference (ROCLING-1997)*.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *CoLing-ACL 1998*.

Baroni, M. and Kilgariff, A. (2006). Large linguistically-processed web corpora for multiple languages. In *EACL 2006*.

Bean, D. (2004). *Acquisition and Application of Contextual Role Knowledge for Coreference Resolution*. Ph.D. thesis, University of Utah, School of Computing.

Bean, D. and Riloff, E. (1999). Corpus-based identification of non-anaphoric noun phrases. In *Proc. ACL 1999*.

Bean, D. and Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *Proc. HLT-NAACL 2004*.

Berger, A. L., Della Pietra, S., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, **22**(1), 39–71.

Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of ACL-1999*.

Beygelzimer, A., Langford, J., and Ravikumar, P. (2008). Multiclass classification with filter trees. (Unpublished Draft, retrieved from `http://hunch.net/~jl/projects/reductions/mc_to_b/invertedTree.pdf`).

Biemann, C. (2002). *Finden von semantischen Relationen in natürlichsprachlichen Texten mit Hilfe maschinellem Lernens*. Diplomarbeit, Universität Leipzig.

Blackburn, P. and Bos, J. (1999). Working with Discourse Representation Theory: An advanced course in computational semantics. `http://homepages.inf.ed.ac.uk/jbos/comsem/book2.html`.

Blaheta, D. and Johnson, M. (2001). Unsupervised learning of multi-word verbs. In *ACL Workshop on Collocations*.

Bos, J., Buitelaar, P., and Mineur, A.-M. (1995). Bridging as coercive accomodation. In S. Manandhar, editor, *Computational Logic for Natural Language Processing (CLNLP '95) - Workshop Proceedings, April 3-5*, South Queensferry, Scotland.

Bosco, C. and Lombardo, V. (2006). Comparing linguistic information in treebank annotations. In *LREC 2006*.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proc. TLT 2002*.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Applied NLP Conference (ANLP-2000)*.

Briscoe, T. and Caroll, J. (2002). Robust accurate statistical annotation of general text. In *LREC 2002*.

Broscheit, S., Ponzetto, S. P., Versley, Y., and Poesio, M. (2010). Extending BART to provide a coreference resolution system for German. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

Brown, R. D. (2002). Corpus-driven splitting of compound words. In *Proceedings of the Ninth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*.

Buchholz, S. and Marsi, E. (2006). Conll-x shared task on multilingual dependency parsing. In *CoNLL 2006*.

Budanitsky, A. and Hirst, G. (2001). Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *Proceedings of the NAACL-01 workshop on WordNet and other lexical resources*.

Buitelaar, P., Declerck, T., Sacaleanu, B., Vintar, S., Raileanu, D., and Crispi, C. (2003). A multi-layered, XML-based approach to the integration of linguistic and semantic annotations. In *EACL 2003 Workshop on Language Technology and the Semantic Web (NLPXML'03)*.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology learning from text: An overview. In *Ontology Learning from Text: Methods, Evaluation and Applications*. IOS Press, Amsterdam.

Bunescu, R. (2003). Associative anaphora resolution: A web-based approach. In *EACL 2003 Workshop on the Computational Treatment of Anaphora*.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2006). The salsa corpus: a german corpus resource for lexical semantics. In *Proc. LREC 2006*.

Burger, J. D. and Conolly, D. (1992). Probabilistic resolution of anaphoric reference. In *Proc. AAAI Fall Symposium on Intelligent Probabilistic Approaches to Language*.

Burnard, L., editor (1995). *Users Reference Guide British National Corpus Version 1.0.* Oxford University Computing Service.

Butt, M., Dipper, S., Frank, A., and King, T. H. (1999). Writing large-scale parallel grammars for english, french, and german. In *Proceedings of the LFG99 Conference*.

Byron, D. (2004). *Resolving Pronominal Reference to Abstract Entities*. Ph.D. thesis, University of Rochester, Department of Computer Science. Technical Report 815.

Byron, D. K. (2001). The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, **27**(4), 569–577.

Byron, D. K. and Gegg-Harrison, W. (2004). Eliminating non-referring noun phrases from coreference resolution. In *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2004)*.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the role of bleu in machine translation research. In *EACL 2006*.

Caraballo, S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*.

Carreras, X., Màrquez, L., and Padró, L. (2003). A simple named entity extractor using adaboost. In *CoNLL 2003 Shared Task*.

Carroll, G. and Rooth, M. (1998). Valence induction with a head-lexicalized PCFG. In *Proceedings fo the Conference for Empirical Methods in Natural Language Processing*.

Carter, D. M. (1985). Common sense inference in a focus-guided anaphor resolver. *Journal of Semantics*, **4**, 237–246.

Castaño, J., Zhang, J., and Pustejovsky, J. (2002). Anaphora resolution in biomedical literature. In *International Symposium on Reference Resolution*.

Cederberg, S. and Widdows, D. (2003). Using LSA and noun coordination information to improve the precision and recall of automatic hyponymy extraction. In *Seventh Conference on Computational Natural Language Learning (CoNLL-2003)*.

Charniak, E. (1972). *Toward a Model of Children's Story Comprehension*. Ph.D. thesis, MIT Computer Science and Artificial Intelligence Lab (CSAIL).

Charniak, E. (1983). Passing markers: A theory of contextual influence in language comprehension. *Cognitive Science*, **7**, 171–190.

Charniak, E. (1996). Tree-bank grammars. Technical Report CS-96-02, Department of Computer Science, Brown University.

Charniak, E. (1997). Statistical parsing with a context-free grammar and word statistics. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI'97)*.

Charniak, E. (2000). A maximum-entropy-inspired parser. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*.

Charniak, E. and Johnson, M. (2005). Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proc. ACL 2005*.

Cheung, J. C. K. and Penn, G. (2009). Topological field parsing of german. In *ACL 2009*.

Chinchor, N. A. (1998). Overview of MUC-7/MET-2. In *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

Chomsky, N. (1965). *Aspects of a theory of Syntax*. MIT Press.

Chomsky, N. (1981). *Lectures on Government and Bindung: The Pisa Lectures*. Foris, Dordrecht.

Christophersen, P. (1939). *The Articles. A Study of Their Theory and Use in English*. Munksgaard.

Church, K. and Gale, W. (1991). Concordances for parallel text. In *Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research*.

Church, K. W. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1), 22–29.

Cimiano, P. and Staab, S. (2004). Learning by googling. *SIGKDD Explorations*, **6**(2), 24–33.

Cimiano, P. and Wenderoth, J. (2005). Automatically learning qualia structures from the web. In *Proceedings of the ACL'05 Workshop on Deep Lexical Acquisition*.

Clark, H. H. (1975). Bridging. In R. C. Schank and B. L. Nash-Webber, editors, *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174, Cambridge, MA. Association for Computing Machinery.

Collins, M. (1997). Three generative, lexicalized models for statistical parsing. In *Proc. ACL 1997*.

Collins, M. (1999). *Head-driven statistical models for natural-language parsing*. Ph.D. thesis, Univ. of Pennsylvania.

Collins, M. (2000). Discriminative reranking for natural language parsing. In *ICML 2000*.

Collins, M. and Koo, T. (2005). Discriminative reranking for natural language parsing. *Computational Linguistics*, **31**(1), 25–69.

Connolly, D., Burger, J. D., and Day, D. S. (1994). A machine learning approach to anaphoric reference. In *Proc. International Conference on New Methods in Language Processing (NeMLaP) 1994*.

Consten, M. and Knees, M. (2005). Complex anaphors - ontology and resolution. In *Proceedings of the 15th Amsterdam Colloquium*.

Culotta, A., Wick, M., and McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *Proc. HLT/NAACL 2007*.

Curran, J. and Moens, M. (2002a). Improvements in automatic thesaurus extraction. In *Proceedings of the Workshop on Unsupervised Lexical Acquisition (SIGLEX)*.

Curran, J. and Moens, M. (2002b). Scaling context space. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Curran, J. R. (2004). *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Dagan, I. and Itai, A. (1990). Automatic processing of large corpora for the resolution of anaphora references. In *CoLing 1990*.

Daum, M., Foth, K., and Menzel, W. (2004). Automatic transformation of phrase treebanks to dependency trees. In *Proc. 4th Int. Conference on Language Resources and Evaluation (LREC 2004)*.

Daumé III, H. and Marcu, D. (2005). A large-scale exploration of effective global features for a joint entity detection and tracking model. In *HLT/EMNLP'05*, pages 97–104.

Deerwester, S., Dumais, S. T., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the Society for Information Science*, **41**(6), 391–407.

Dempster, A. (1968). A generalization of bayesian inference. *Journal of the Royal Statistical Society Series B*, **30**, 205–247.

Denis, P. and Baldridge, J. (2007a). Joint determination of anaphoricity and coreference resolution using integer programming. In *NAACL 2007*.

Denis, P. and Baldridge, J. (2007b). A ranking approach to pronoun resolution. In *Proc. IJCAI 2007*.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.

Drach, E. (1937). *Grundgedanken der Deutschen Satzlehre*. Diesterweg.

Dubey, A. (2005). What to do when lexicalization fails: parsing German with suffix analysis and smoothing. In *ACL-2005*.

Dubey, A. and Keller, F. (2003). Probabilistic parsing for German using sister-head dependencies. In *ACL'2003*.

Dumais, S. (1991). Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, **23**(2), 229–236.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., and Harshman, R. (1988). Using latent semantic analysis to improve information retrieval. In *Proceedings of CHI'88: Conference on Human Factors in Computing*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.

Eisner, J. and Satta, G. (1999). Efficient parsing for bilexical context-free grammars and head automaton grammars. In *ACL 1999*.

Erdmann, O. (1886). *Grundzüge der deutschen Syntax nach ihrer geschichtlichen Entwicklung dargestellt*. Verlag der Cotta'schen Buchhandlung, Stuttgart.

Evans, R. (2003). A framework for named entity recognition in the open domain. In *RANLP 2003*.

Evert, S. (2005). *The Statistics of Word Cooccurrences*. Ph.D. thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart.

Fauconnier, G. (1984). *Espaces Mentaux*. Editions de Minuit, Paris.

Finley, T. and Joachims, T. (2005). Supervised clustering with support vector machines. In *Proc. International Conference on Machine Learning (ICML)*.

Fisher, D., Soderland, S., McCarthy, J., Feng, F., and Lehnert, W. (1996). Description of the UMass system as used for MUC-6. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*.

Fleischman, M., Hovy, E., and Echihabi, A. (2003). Offline strategies for online question answering: Answering questions before they are asked. In *Proc. ACL 2003*.

Fligelstone, S. (1992). Developing a scheme for annotating text to show anaphoric relations. In *New Directions in English Language Corpora: Methodology, Results, Software Development*, pages 153–170. Mouton de Gruyter, Berlin.

Forst, M. (2003). Treebank conversion - creating an f-structure bank from the TIGER corpus. In *Procedings of the LFG03 conference*.

Foth, K. (2003). Eine umfassende Dependenzgrammatik des Deutschen. Technical report, Fachbereich Informatik, Universität Hamburg.

Foth, K. and Menzel, W. (2003). Subtree parsing to speed up deep analysis. In *Proc. 8th Int. Workshop on Parsing Technologies (IWPT-2003)*.

Foth, K. and Menzel, W. (2006). Hybrid parsing: Using probabilistic models as predictors for a symbolic parser. In *ACL 2006*.

Foth, K., Daum, M., and Menzel, W. (2004). A broad-coverage parser for German based on defeasible constraints. In *KONVENS 2004, Beiträge zur 7. Konferenz zur Verarbeitung natürlicher Sprache*, pages 45–52, Wien.

Fraurud, K. (1996). Cognitive ontology and np form. In Fretheim and Gundel, editors, *Reference and Referent Accessibility*, pages 65–87. John Benjamins.

Frege, G. (1879). *Begriffsschrift: Eine der arithmetischen nachgebildeten Formelsprache des reinen Denkens*. Neubert, Halle.

Gardent, C. and Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des déscriptions d éfinies. *Traitement Automatique des Langues*, **46**(1), 115–140.

Garera, N. and Yarowsky, D. (2006). Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *Proc. CoNLL 2006*.

Gasperin, C. and Vieira, R. (2004). Using word similarity lists for resolving indirect anaphora. In *ACL'04 workshop on reference resolution and its applications*.

Gasperin, C., Gamallo, P., Augustini, A., Lopes, G., and de Lima, V. (2001). Using syntactic contexts for measuring word similarity. In *ESSLLI 2001 Workshop on Knowledge Acquisition and Categorization*.

Gasperin, C., Salmon-Alt, S., and Vieira, R. (2004). How useful are similarity word lists for indirect anaphora resolution? In *Proc. DAARC 2004*.

Geach, P. (1967). Intentional identity. *Journal of Philosophy*, **74**(20), 627–632.

Geffet, M. and Dagan, I. (2004). Feature vector quality and distributional similarity. In *CoLing 2004*.

Geleijnse, G. and Korst, J. (2006). Learning effective surface text patterns for information extraction. In *Proceedings of the EACL 2006 workshop on Adaptive Text Extraction and Mining (ATEM 2006)*.

Gentile, C. (2001). A new approximate maximal margin classification algorithm. *Journal of Machine Learning Research*, **2**, 213–242.

Giesbrecht, E. (2008). *An Evaluation of POS Taggers for the Web as Corpus*. Master's thesis, Universität Osnabrück.

Girju, R. (2003). Automatic detection of causal relations for question answering. In *ACL 2003 Workshop on "Multilingual Summarization and Question Answering - Machine Learning and Beyond"*.

Goecke, D., Stührenberg, M., and Wandmacher, T. (2008). A hybrid approach to resolve nominal anaphora. *LDV Forum - Zeitschrift für Computerlinguistik und Sprachtechnologie*, **23**(1), 43–58.

Grefenstette, G. (1992). Use of syntactic context to produce term association lists for text retrieval. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 89–97.

Grishman, R. and Sundheim, B. (1995). Design of the MUC-6 evaluation. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

Groenendijk, J. and Stokhof, M. (1984). *Studies on the semantics of questions and the pragmatics of answers*. Ph.D. thesis, Universtiteit van Amsterdam.

Gundel, J. K., Hedberg, N., and Zacharsky, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, **69**, 247–307.

Haapalainen, M. and Majorin, A. (1995). GERTWOL und Morphologische Disambiguierung fürs Deutsche. In *NODALIDA 1995*.

Haghigi, A. and Klein, D. (2007). Unsupervised coreference resolution in a nonparametric bayesian model. In *ACL-2007*.

Hall, J. and Nivre, J. (2008). Parsing discontinuous phrase structure with grammatical functions. In *Proceedings of the 6th International Conference on Natural Language Processing (GoTAL 2008)*.

Hall, J., Nivre, J., and Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Harabagiu, S., Bunescu, R., and Maiorano, S. (2001). Text and knowledge mining for coreference resolution. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001)*.

Harabagiu, S. M., Miller, G., and Moldovan, D. (1999). Wordnet 2 - a morphologically and semantically enhanced resource. In *SIGLEX 1999*.

Hartrumpf, S. (2001). Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proceedings of the Fifth Conference on Natural Language Learning (CoNLL-2001)*.

Hartrumpf, S., Helbig, H., and Osswald, R. (2003). The semantically based corpus HaGenLex - structure and technological environment. *Traitement automatique des langues*, **44**(2), 81–105.

Hawkins, J. (1978). *Definiteness and Indefiniteness*. Croom Helm, London.

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of the 14th International Conference on Computational Linguistics (COLING 92)*.

Hearst, M. (1998). Automated discovery of wordnet relations. In *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge (MA), USA.

Heim, I. (1982). *The Semantics of Definite and Infedinite Noun Phrases*. Ph.D. thesis, University of Massachusetts at Amherst.

Helbig, H. (2006). *Knowledge Representation and the Semantics of Natural Language*. Springer.

Helbig, H. and Hartrumpf, S. (1997). Word class functions for syntactic-semantic analysis. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Vloet, J. V. D., and Verschelde, J.-L. (2008). A coreference corpus and resolution system for dutch. In *LREC 2008*.

Hengel, C. and Pfeifer, B. (2005). Kooperation der Personennamendatei (PND) mit Wikipedia. *Dialog mit Bibliotheken*, **17**(3), 18–24.

Hindle, D. (1983). User manual for Fidditch. Technical memorandum 7590–142, Naval Research Laboratory.

Hindle, D. (1990). Noun classification from predicate argument structures. In *Proceedings of the 28th annual meeting of the Association for Computational Linguistics*.

Hinrichs, E. and Nakazawa, T. (1989). Flipped out: AUX in German. In *Papers from the 25th Annual Regional Meeting of the Chicago Linguistic Society*.

Hinrichs, E., Kübler, S., and Naumann, K. (2005a). A unified representation for morphological, syntactic, semantic and referential annotations. In *ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, Ann Arbor.

Hinrichs, E., Filippova, K., and Wunsch, H. (2005b). What treebanks can do for you: Rule-based and machine-learning approaches to anaphora resolution in German. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT'05)*.

Hinrichs, E. W., Filippova, K., and Wunsch, H. (2005c). A data-driven approach to pronominal anaphora resolution in german. In *RANLP 2005*.

Hirschman, L. (1992). An adjunct test for discourse processing in MUC-4. In *Proceedings of the 4th conference on Message understanding (MUC-4)*.

Hirschman, L. and Chinchor, N. (1997). MUC-7 coreference task definition (version 3.0). In *Proceedings of the 7th Message Understanding Conference*.

Hobbs, J. (1978). Resolving pronoun references. *Lingua*, **44**, 311–338.

Hobbs, J. (1985). Granularity. In *Proc. Ninth International Joint Conference on Artificial Intelligence (IJCAI 1985)*.

Hobbs, J., Appelt, D., Bear, J., Israel, D., Kameyama, M., Stickel, M., and Tyson, M. (1996). FASTUS: A cascaded finite-state transducer for extracting information from natural language text. In E. Roche and Y. Schabes, editors, *Finite State Devices for Natural Language Processing*. MIT Press.

Holler-Feldhaus, A. (2004). Koreferenz in Hypertexten: Anforderungen an die Annotation. *Osnabrücker Beiträge zur Sprachtheorie (OBST)*, **68**, 9–29.

Hollingshead, K., Fisher, S., and Roark, B. (2005). Comparing and combining finite-state and context-free parsers. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*.

Huang, L. (2008). Forest reranking: Discriminative parsing with non-local features. In *HLT/ACL 2008*.

Hudson, R. (1984). *Word Grammar*. Blackwell, Oxford.

Humphreys, K., Gaizauskas, R., Azzam, S., Huyck, C., Mitchell, B., Cunningham, H., and Wilks, Y. (1998). University of sheffield: Description of the lasie-ii system as used for muc-7. In *Proceedings of MUC-7*.

Höhle, T. (1986). Der Begriff "Mittelfeld", Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, pages 329–340.

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a japanese text corpus with predicate-argument and coreference relations. In *ACL'07 Linguistic Annotation Workshop (LAW)*.

Jackendoff, R. (2002). *Foundations of language: brain, meaning, grammar, evolution*. Oxford University Press, Oxford.

Ji, H., Westbrook, D., and Grishman, R. (2005). Using semantic relations to refine coreference decisions. In *HLT-EMNLP'2005*.

Johnson, M. (2001). Trading recall for precision with confidence sets. Technical report, Brown University.

Johnson, M., Geman, S., Canon, S., Chi, Z., and Riezler, S. (1999). Estimators for stochastic "unification-based" grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-1999)*.

Kabadjov, M. A. (2007). *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification*. unpublished PhD thesis, Department of Computer Science, University of Essex.

Kadmon, N. (1990). Uniqueness. *Linguistics and Philosophy*, **13**(3), 273–324.

Kameyama, M. (1997). Recognizing referential links: an information extraction prespective. In *ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*. Mathematisch Centrum, Amsterdam.

Kamp, H. and Reyle, U. (1993). *From Discourse to Logic*. Kluwer, Dordrecht.

Karttunen, L. (1976). Discourse referents. In J. D. McCawley, editor, *Syntax and Semantics 7: Notes from the Linguistic Underground*, pages 363–385. Academic Press.

Katrenko, S. and Adriaans, P. (2008). Qualia structures and their impact on the concrete noun categorization task. In *ESSLLI 2008 workshop on Distributional Lexical Semantics*.

Kaup, B., Kelter, S., and Habel, C. (1999). Taking the functional aspect of mental models as a starting point for studying discourse comprehension. In G. Rickheit and C. Habel, editors, *Mental Models in Discourse Processing and Reasoning*, volume 128 of *Advances in Psychology*. North Holland, Amsterdam.

Kehler, A., Appelt, D., Taylor, L., and Simma, A. (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pages 289–296.

Kermes, H. and Evert, S. (2002). Yac - a recursive chunker for unrestricted German text. In *Proc. LREC 2002*.

Kibble, R. and van Deemter, K. (2000). Coreference annotation: Whither? In *LREC 2000*.

Kilgrariff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, **29**(3), 333–348.

Kintsch, W. and van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, **85**, 363–394.

Klein, D. and Manning, C. D. (2003a). Accurate unlexicalized parsing. In *ACL 2003*.

Klein, D. and Manning, C. D. (2003b). Fast exact inference with a factored model for natural language parsing. In *NIPS 2002*.

Klenner, M. (2007). Enforcing consistency of coreference sets. In *RANLP 2007*.

Klenner, M. and Ailloud, E. (2008). Enhancing coreference clustering. In *Second Bergen Workshop on Anaphora Resolution (WAR II)*.

Knees, M. (2006). The German temporal anaphor danach - Ambiguity in interpretation and annotation. In *ESSLLI 2006 workshop on Ambiguity and Anaphora*.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

Koehn, P. and Knight, K. (2003). Empirical methods for compound splitting. In *EACL 2003*.

Kouchnir, B. (2004). A machine learning approach to German pronoun resolution. In *Proceedings of the ACL'04 Student Research Workshop*.

Krahmer, E. (1998). *Presupposition and Anaphora*. CSLI Publications, Stanford, CA.

Krahmer, E. and van Deemter, K. (1998). On the interpretation of anaphoric noun phrases: Towards a full understanding of partial matches. *Journal of Semantics*, **15**(2), 355–392.

Krifka, M. (1987). An outline of genericity. SNS-Bericht 87-25, Seminar für natürlich-sprachliche Systeme, Universität Tübingen.

Kudo, T. and Matsumoto, Y. (2004). A boosting algorithm for classification of semi-structured text. In *EMNLP 2004*.

Kunz, K. and Hansen-Schirra, S. (2003). Coreference annotation of the tiger treebank. In *TLT 2003*.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of LREC 2002*.

Kupietz, M. and Keibel, H. (2009). The Mannheim German reference corpus (DeReKo) as a basis for empirical linguistic research. *Working Papers in Corpus-Based Linguistics and Language Education*, **3**, 53–59.

Kübler, S. (2008). The PaGe 2008 shared task on parsing German. In *Proceedings of the ACL-2008 Workshop on Parsing German*.

Kübler, S. and Telljohann, H. (2002). Towards a dependency-based evaluation for partial parsing. In *LREC 2002*.

Kübler, S., Hinrichs, E. W., and Maier, W. (2006). Is it really that difficult to parse German? In *EMNLP'06*.

Kübler, S., Maier, W., Rehbein, I., and Versley, Y. (2008). How to compare treebanks. In *LREC 2008*.

Kübler, S., Hinrichs, E., Maier, W., and Klett, E. (2009). Parsing coordinations. In *EACL 2009*.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition. *Psychological Review*, **104**(2), 211–240.

Lappin, S. and Leass, H. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4), 535–561.

Larson, M., Willett, D., Köhler, J., and Rigoll, G. (2000). Compound splitting and lexical unit recombination for improved performance of a speech recognition system for german parliamentary speeches. In *6th Int. Conference on Spoken Language Processing (ICSLP)*.

Lasnik, H. (1989). *Essays on Anaphora*. Kluwer, Dordrecht.

Lavelli, A. and Corazza, A. (2009). The Berkeley Parser at the EVALITA constituency parsing task. In *Proceedings of the Workshop on Evaluation of NLP Tools for Italian (EVALITA 2009)*.

Lee, L. (1999). Measures of distributional similarity. In *ACL 1999*.

Lee, L. (2001). On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*.

Lehnert, W., Cardie, C., McCarthy, J., Riloff, E., and Soderland, S. (1992). University of massachusetts: Description of the CIRCUS system as used for MUC-4. In *Proceedings of the fourth Message Understanding Conference (MUC-4)*.

Leidner, J. (2007). *Toponym Resolution in Text*. Ph.D. thesis, Institute for Communicating and Collaborative Systems, University of Edinburgh.

Lemnitzer, L. and Zinsmeister, H. (2006). *Einführung in die Korpuslinguistik*. Narr, Tübingen.

Levy, R. and Manning, C. (2004). Deep dependencies from context-free statistical parsers: correcting the surface dependency approximation. In *ACL 2004*.

Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic*, **8**, 339–359.

Lezius, W., Dipper, S., and Fitschen, A. (2000). IMSLex - representing morphological and syntactic information in a relational database. In U. Heid, S. Evert, E. Lehmann, and C. Rohrer, editors, *Proceedings of the 9th EURALEX International Congress*, pages 133–139.

Lin, D. (1995). University of Manitoba: Description of the PIE system used for MUC-6. In *Proceedings of the 6th Message Understanding Conference*.

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proc. CoLing/ACL 1998*.

Lin, D. (1998b). Dependency-based evaluation of Minipar. In *Workshop on the Evaluation of Parsing Systems*.

Lin, D. (1998c). An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*.

Lin, D. (1998d). Using collocation statistics in information extraction. In *Proceedings of MUC-7*.

Link, G. (1983). The logical analysis of plurals and mass terms: a lattice-theoretical approach. In R. Bäuerle, C. Schwarze, and A. von Stechow, editors, *Meaning, Use and Interpretation of Language*. de Gruyter, Berlin.

Liu, D. C. and Nocedal, J. (1989). On the limited memory method for large scale optimization. *Mathematical Programming B*, **45**(3), 503–528.

Lund, K., Atchley, R. A., and Burgess, C. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proc. of the 17th Annual Conference of the Cognitive Science Society*, pages 660–665.

Luo, X. (2005). On coreference resolution performance metrics. In *HLT-EMNLP 2005*.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. (2004). A mention-synchronous coreference resolution algorithm based on the Bell tree. In *ACL 2004*.

Löbner, S. (1985). Definites. *Journal of Semantics*, **4**, 279–326.

Magerman, D. M. (1995). Statistical decision-tree models for parsing. In *ACL'1995*.

Magnini, B., Pianta, E., Girardi, C., Negri, M., Romano, L., Speranza, M., Lenzi, V. B., and Sprugnoli, R. (2006). I-CAB: the Italian Content Annotation Bank. In *Proc. LREC 2006*.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*.

Mani, I. (1998). A theory of granularity and its application to problems of polysemy and underspecification of meaning. In A. G. Cohn, L. K. Schubert, and S. C. Shapiro, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixth Internatinal Conference (KR'98)*, pages 245–255, San Mateo. Morgan Kaufmann.

Mann, G. S. and Yarowsky, D. (2003). Unsupervised personal name disambiguation. In *CoNLL 2003*.

Manning, C. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marek, T. (2006). *Analysis of German Compounds Using Weighted Finite State Transducers*. Bachelor of arts thesis, Universität Tübingen.

Markert, K. and Nissim, M. (2005). Comparing knowledge sources for nominal anaphora resolution. *Computational Linguistics*, **31**(3), 367–402.

Markert, K., Nissim, M., and Modjeska, N. N. (2003). Using the web for nominal anaphora resolution. In *EACL Workshop on the computational treatment of anaphora*.

Matsuzaki, T., Miyao, Y., and Tsujii, J. (2005). Probabilistic CFG with latent annotations. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding predominant word senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*.

McCarthy, J. F. (1996). *A Trainable Approach to Coreference Resolution for Information Extraction*. Ph.D. thesis, University of Massachusetts.

McCarthy, J. F. and Lehnert, W. G. (1995). Using decision trees for coreference resolution. In *Proc. IJCAI 1995*.

McDonald, R. (2006). Online learning of approximate dependency parsing algorithms. In *EACL 2006*.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of HLT/EMNLP 2005*.

McDonald, S. and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proc. 23rd Annual Conference of the Cognitive Society*.

McEnery, T. and Wilson, A. (1996). *Corpus Linguistics*. Edinburgh University Press.

Mengel, A., Dybkjaer, L., Garrido, J., Heid, U., Klein, M., Pirrelli, V., Poesio, M., Quazza, S., Schiffrin, A., and Soria, C. (2000). Mate dialogue annotation guidelines. MATE Deliverable D2.1; `http://www.ims.uni-stuttgart.de/projekte/mate/mdag/`.

Meyer, J. and Dale, R. (2002). Using the wordnet hierarchy for associative anaphora resolution. In *Proceedings of SemaNet'02: Building and Using Semantic Networks*.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, **6**(1), 1–28.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of English. *Cognition*, **41**, 197–229.

Mitkov, R. (1998). Robust pronoun resolution with limited knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (Coling-ACL 1998)*.

Mitkov, R. (1999). Anaphora resolution: The state of the art. Working paper (based on the COLING/ACL 1998 tutorial on anaphora resolution), University of Wolverhampton.

Mitkov, R. (2000). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In *DAARC 2000*.

Mitkov, R. (2005). Anaphora resolution. In *Oxford Handbook of Computational Linguistics*. Oxford University Press.

Miyao, Y., Sætre, R., Sagae, K., Matsuzaki, T., and Tsujii, J. (2008). Task-oriented evaluation of syntactic parsers and their representations. In *ACL 2008*.

Modjeska, N. N., Markert, K., and Nissim, M. (2003). Using the web in machine learning for other-anaphora resolution. In *EMNLP 2003*.

Montague, R. (1973). The proper treatment of quantification in ordinary English. In K. J. J. Hintikka, J. M. E. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language*, pages 221–242. Reidel, Dordrecht.

Monz, C. and de Rijke, M. (2001). Shallow morphological analysis in monolingual information retrieval for dutch, german and italian. In *CLEF 2001*.

Morton, T. S. (2000). Coreference for NLP applications. In *Proc. ACL 2000*.

Müller, C., Rapp, S., and Strube, M. (2002). Applying co-training to reference resolution. In *ACL-02*.

Müller, F. H. (2004a). Annotating grammatical functions in German using finite-state cascades. In *Proc. 20th Int. Conference on Computational Linguistics (COLING 2004)*.

Müller, F. H. (2004b). Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). Technischer bericht, Seminar für Sprachwissenschaft, Universität Tübingen.

Müller, F. H. and Ule, T. (2002). Annotating topological fields and chunks – and revising POS tags at the same time. In *Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002)*.

Nakov, P., Angelova, G., and von Hahn, W. (2002). Automatic recognition and morphological classification of unknown German nouns. Bericht 243, Universität Hamburg, Fachbereich Informatik. `http://nats-www.informatik.uni-hamburg.de/~vhahn/Downloads/Report243.pdf`.

Naumann, K. (2006). Manual for the annotation of in-document referential relations. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Ng, V. (2007). Shallow semantics for coreference resolution. In *Proc. IJCAI 2007*.

Ng, V. and Cardie, C. (2002a). Combining sample selection and error-driven pruning for machine learning of coreference rules. In *Coreference on Empirical Methods in Natural Language Processing (EMNLP)*.

Ng, V. and Cardie, C. (2002b). Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING 2002*.

Ng, V. and Cardie, C. (2002c). Improving machine learning approaches to coreference resolution. In *40th Annual Meeting of the Asssociation for Computational Linguistics*.

Niessen, S. and Ney, H. (2000). Improving SMT quality with morpho-syntactic analysis. In *Coling 2000*.

Nivre, J. (2003). An efficient algorithm for projective dependency parsing. In *8th International Workshop on Parsing Technologies*.

Padó, S. and Lapata, M. (2003). Constructing semantic space models from parsed corpora. In *Proceedings of ACL 2003*.

Padó, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, **33**(2), 161–199.

Partee, B. H. (1978). Bound variables and other anaphors. In D. L. Waltz, editor, *Theoretical Issues in Natural Language Processing 2 (TINLAP-2)*, Urbana, IL (USA). University of Illinois.

Passonneau, R. (1996). Instructions for applying discourse reference annotation for multiple applications (DRAMA). Unpublished manuscript.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet::similarity - measuring the relatedness of concepts. In *NAACL 2004 demo*.

Peterson, P., Martins, J. R. R. A., and Alonso, J. J. (2001). Fortran to Python interface generator with an application to aerospace engineering. In *Proceedings of the 9th International Python Conference, Long Beach, CA*.

Petrov, S. and Klein, D. (2007). Improved inference for unlexicalized parsing. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*.

Petrov, S. and Klein, D. (2008a). Discriminative log-linear grammars with latent variables. In *NIPS 2008*.

Petrov, S. and Klein, D. (2008b). Parsing German with latent variable grammars. In *Parsing German Workshop at ACL-HLT 2008*.

Petrov, S., Barett, L., Thibaux, R., and Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. In *COLING-ACL 2006*.

Philips, M. (1985). *Aspects of Text Structure: An investigation of the lexical organization of text*. Elsevier, Amsterdam.

Pinkal, M. (2002). Sehr große Korpora für große Wörterbücher. Vortragsfolien zum "Kolloquium Korpus-Annotierung". http://www.uni-saarland.de/fak4/fr46/steiner/dandelion/kolloq-ldb-fe02/pinkal.pdf.

Poesio, M. (1994). Weak definites. In *Proceedings of the Fourth Conference on Semantics and Linguistic Theory (SALT-4)*.

Poesio, M. (2004). The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proc. of SIGDIAL'04*, Boston.

Poesio, M. and Artstein, R. (2008). Anaphoric annotation in the ARRAU corpus. In *LREC 2008*.

Poesio, M. and Reyle, U. (2001). Underspecification in anaphoric reference. In *Fourth International Workshop on Computational Semantics (IWCS-4)*.

Poesio, M., Vieira, R., and Teufel, S. (1997). Resolving bridging descriptions in unrestricted text. In *ACL-97 Workshop on Operational Factors in Practical, Robust, Anaphora Resolution For Unrestricted Texts*.

Poesio, M., Schulte im Walde, S., and Brew, C. (1998). Lexical clustering and definite description interpretation. In *AAAI Spring Symposium on Learning for Discourse*.

Poesio, M., Henschel, R., Hitzeman, J., Kibble, R., Montague, S., and van Deemter, K. (1999). Towards an annotation scheme for noun phrase generation. In *Proceedings of the EACL workshop on Linguistically Interpreted Corpora (LINC-99)*.

Poesio, M., Ishikawa, T., Schulte im Walde, S., and Vieira, R. (2002). Acquiring lexical knowledge for anaphora resolution. In *Language resources and evaluation conference (LREC 2002)*.

Poesio, M., Reyle, U., and Stevenson, R. (2003). Justified sloppiness in anaphoric reference. In H.Bunt and R.Muskens, editors, *Computing Meaning 3*. Kluwer. to appear.

Poesio, M., Mehta, R., Maroudas, A., and Hitzeman, J. (2004a). Learning to resolve bridging references. In *ACL-2004*.

Poesio, M., Delmonte, R., Bristot, A., Chiran, L., and Tonelli, S. (2004b). The VENEX corpus of anaphora and deixis in spoken and written Italian. Available

at http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf.

Poesio, M., Alexandrov-Kabadjov, M., Vieira, R., Goulart, R., and Uryupina, O. (2005). Does discourse-new detection help definite description resolution? In *Proceedings of the 6th International Workshop on Computational Semantics (IWCS-6)*.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. HLT/NAACL 2006*.

Ponzetto, S. P. and Strube, M. (2007). Deriving a large-scale taxonomy from wikipedia. In *AAAI 2007*.

Pradhan, S., Ramshaw, L., Weischedel, R., MacBride, J., and Micciulla, L. (2007). Unrestricted coreference: Identifying entities and events in ontonotes. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*.

Prescher, D. (2005). Inducing head-driven PCFGs with latent heads: Refining a tree-bank grammar for parsing. In *Proceedings of the 16th European Conference on Machine Learning (ECML 2005)*.

Prince, E. F. (1992). The ZPG letter: subjects, definiteness and information-status. In S. Thompson and W. Mann, editors, *Discourse description: diverse analyses of a fund raising text*. John Benjamins B.V., Amsterdam.

Rafferty, A. and Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *ACL'08 workshop on Parsing German*.

Ramshaw, L. and Marcus, M. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proc. Ninth Machine Translation Summit*.

Ravichandran, D., Pantel, P., and Hovy, E. (2005). Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proc. ACL 2005*.

Reboul, A. (1997). What (if anything) is accessibility? a relevance-oriented criticism of Ariel's accessibility theory of referring expressions. In J. H. Connolly, R. M. Vismans, C. S. Butler, and R. A. Gatwards, editors, *Discourse and pragmatics in functional grammar*. Mouton de Gruyter, Berlin.

Recasens, M. and Marti, M. A. (2009). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, **to appear**, 1–31.

Recasens, M., Martí, M., and Taulé, M. (2007). Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE corpus. In *RANLP 2007*.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. (2010). Semeval task 1: Coreference resolution in multiple languages. In *Proceedings of the ACL 2010 Workshop on Semantic Evaluations (SemEval 2010)*. To appear.

Reinhart, T. (1976). *The Syntactic Domain of Anaphora*. Ph.D. thesis, Massachusetts Institute of Technology.

Reinhart, T. (1983). *Anaphora and Semantic Interpretation*. Croom Helm, London.

Reitsma, F. and Bittner, T. (2003). Process, hierarchy and scale. In W. Kuhn, M. Worboys, and S. Timpf, editors, *Spatial Information Theory. Cognitive and Computational Foundations of Geographic Information Science (COSIT'03)*.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a semantically annotated lexicon via EM-based clustering. In *ACL 1999*.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The Reuters corpus volume 1 - from yesterday's news to tomorrow's language resources. In *LREC 2002*.

Rosenfeld, R. (1996). A maximum entropy approach to adaptive statistical language modeling. *Computer Speech and Language*, **10**, 187–228.

Russell, B. (1905). On denoting. *Mind*, **14**(56), 479–490.

Rössler, M. (2006). *Korpus-adaptive Eigennamenerkennung*. Ph.D. thesis, University of Duisburg-Essen.

Sanford, A. J. and Moxey, L. M. (1999). What are mental models made of? In G. Rickheit and C. Habel, editors, *Mental Models in Discourse Processing and Reasoning*, volume 128 of *Advances in Psychology*. North Holland, Amsterdam.

Schiehlen, M. (2004a). Annotation strategies for probabilistic parsing in German. In *Proc. Coling 2004*.

Schiehlen, M. (2004b). Optimizing algorithms for pronoun resolution. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 390–396, Geneva, Switzerland.

Schiller, A. (2005). German compound analysis with *wfsc*. In *Proc. FSMNLP 2005*.

Schiller, A. and Stöckert, C. (1995). DMOR. Technical report, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schiller, A., Teufel, S., and Thielen, C. (1995). Guidelines für das Tagging deutscher Texte mit STTS. Technical report, Universität Stuttgart / Universität Tübingen.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proc. ACL-SIGDAT Workshop*.

Schmid, H. (2004). Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proc. Coling 2004*.

Schmid, H. (2006). Trace prediction and recovery with unlexicalized PCFGs and slash features. In *Proceedings of COLING-ACL 2006*.

Schmid, H. and Schulte im Walde, S. (2000). Robust German noun chunking with a probabilistic context-free grammar. In *Proc. COLING 2000*.

Schmid, H., Fitschen, A., and Heid, U. (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proceedings of LREC 2004*.

Schulte im Walde, S. (2000). The German statistical grammar model: Development, training and linguistic exploitation. Arbeitspapiere des Sonderforschungsbereich 340 *Linguistic Theory and the Foundations of Computational Linguistics* 162, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schulte im Walde, S. (2006). Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, **32**, 159–194.

Schwarz, M. (2000). *Indirekte Anaphern in Texten*, volume 413 of *Linguistische Arbeiten*. Niemeyer, Tübingen.

Schütze, H. (1992). Dimensions of meaning. In *Proc. Supercomputing 1992*.

Seddah, D., Candito, M., and Crabbé, B. (2009). Cross parser evaluation and tagset variation : a french treebank study. In *Proceedings of the 11th International Conference on Parsing Technologies (IWPT 2009)*.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shen, L., Satta, G., and Joshi, A. (2007). Guided learning for bidirectional sequence classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*.

Sidner, C. (1979). *Towards a Computational Theory of Definite Anaphora Comprehension in English Discourse*. Ph.D. thesis, Massachusetts Institute of Technology.

Skut, W. and Brants, T. (1998). A maximum-entropy partial parser for unrestricted text. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

Skut, W., Krenn, B., Brants, T., and Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*.

Smith, B. and Brogaard, B. (2001). A unified theory of truth and reference. *Logique et Analyse*, **43**(169-170), 49–93.

Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In *NIPS 2005*.

Snow, R., Jurafsky, D., and Ng, A. Y. (2006). Semantic taxonomy induction from heterogenous evidence. In *Proceedings of COLING/ACL 2006*.

Soon, W. M., Ng, H. T., and Lim, C. Y. (1999). Corpus-based learning for noun phrase coreference resolution. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-99)*, pages 285–291.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, **27**(4), 521–544.

Spitzer, R. and Fleiss, J. (1974). A re-analysis on the reliability of psychiatric diagnosis. *British Journal on Psychiatry*, **125**, 341–347.

Stede, M. (2004). The Potsdam Commentary Corpus. In *ACL'04 Workshop on Discourse Annotation*.

Stegmann, R., Telljohann, H., and Hinrichs, E. (2000). Stylebook for the German treebank in VERBMOBIL. VM Report 239, Seminar für Sprachwissenschaft, Universität Tübingen.

Steinberger, J., Kabadjov, M., Poesio, M., and Sanchez-Graillet, O. (2005). Improving LSA-based summarization with anaphora resolution. In *HLT/EMNLP'2005*, pages 1–8.

Steinberger, J., Poesio, M., Kabadjov, M., and Jezek, K. (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management*, **43**, 1663–1680. Special issue on Summarization.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *LREC 2006*.

Strawson, P. F. (1950). On referring. *Mind, New Series*, **59**(235), 320–344.

Strube, M., Rapp, S., and Müller, C. (2002). The influence of minimum edit distance on reference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 312–319.

Stuckardt, R. (2001). Design and enhanced evaluation of a robust anaphor resolution algorithm. *Computational Linguistics*, **27**(4), 479–506.

Stuckardt, R. (2003). Coreference-based summarization and question answering: a case for high precision anaphor resolution. In *Symposion on Reference Resolution and Its Application to QA and TS (ARQAS)*.

Stuckardt, R. (2005). Getting started with ROSANA-Deutsch. http://www.stuckardt.de/GettingStarted_ROSANA_Deutsch.pdf.

Telljohann, H., Hinrichs, E. W., and Kübler, S. (2003). Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, Seminar für Sprachwissenschaft, Universität Tübingen.

Tetrault, J. (2001). A corpus-based evaluation of centering and pronoun resolution. *Computational Linguistics*, **27**(4), 507–520.

Thelen, M. and Riloff, E. (2002). A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *EMNLP'2002*.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL 2003*.

Trask, R. L. (1993). *A Dictionary of Grammatical Terms in Linguistics*. Routledge.

Trouilleux, F., Gaussier, E., Bies, G. G., and Zaenen, A. (2000). Coreference resolution evaluation based on descriptive specificity. In *LREC 2000*.

Trushkina, J. (2004). *Morphological disambiguation and dependency parsing of German*. Ph.D. thesis, Seminar für Sprachwissenschaft, Universität Tübingen.

Trushkina, J. and Hinrichs, E. (2004). A hybrid model for morpho-syntactic annotation of german with a large tagset. In D. Lin and D. Wu, editors, *Proceedings of EMNLP 2004*, pages 238–245, Barcelona, Spain. Association for Computational Linguistics.

Ule, T. (2003). Directed treebank refinement for PCFG parsing. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*.

Uryupina, O. (2003). High-precision identification of discourse new and unique noun phrases. In *Proceedings of the ACL Student Workshop*.

Uryupina, O. (2006). Coreference resolution with and without linguistic knowledge. In *Proc. LREC 2006*.

Uryupina, O. (2007). *Knowledge Acquisition for Coreference Resolution*. Ph.D. thesis, Universität des Saarlandes.

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

van der Sandt, R. (1992). Presupposition projection as anaphora resolution. *Journal of Semantics*, **9**(4), 333–377.

van Rijsbergen, C. J. K. (1979). *Information Retrieval*. Butterworths.

Versley, Y. (2005). Parser evaluation across text types. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*.

Versley, Y. (2006). A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.

Versley, Y. (2007). Using the Web to resolve coreferent bridging in German newspaper text. In *Proceedings of GLDV-Frühjahrstagung 2007*, Tübingen. Narr.

Versley, Y. (2008a). Decorrelation and shallow semantic patterns for distributional clustering of nouns and verbs. In M. Baroni, S. Evert, and A. Lenci, editors, *Proceedings of the ESSLLI 2008 Workshop on Distributional Lexical Semantics*.

Versley, Y. (2008b). Vagueness and referential ambiguity in a large-scale annotated corpus. *Research on Language and Computation*, **6**(3–4), 333–353.

Versley, Y. and Rehbein, I. (2009). Scalable discriminative parsing for German. In *Proc. IWPT 2009*.

Vieira, R. and Poesio, M. (1996). Corpus-based approaches to NLP: a practical prototype. In *Anais do XVI Congresso da Sociedade Brasileira de Computação*.

Vieira, R. and Poesio, M. (1997). Processing definite descriptions in corpora. In S. Botley and M. McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*. UCL Press.

Vieira, R. and Poesio, M. (2000). An empirically based system for processing definite descriptions. *Computational Linguistics*, **26**(4), 539–593.

Vieira, R. and Teufel, S. (1997). Towards resolution of bridging descriptions. In *ACL-EACL 1997*.

Vieira, R., Gasperin, C., and Salmon-Alt, S. (2002). Coreference and anaphoric relations of demonstrative noun phrases in multilingual corpus. In *DAARC 2002*.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*. Morgan Kaufmann.

von Heusinger, K. (1997). Definite descriptions and choice functions. In S. Akama, editor, *Logic, Language and Computation*, pages 61–91. Kluwer, Dordrecht.

Wagner, A. and Zeisler, B. (2004). A syntactically annotated corpus of Tibetan. In *LREC 2004*.

Walker, C., Strassel, S., Medero, J., and Maeda, K. (2006). ACE 2005 multilingual training corpus. LDC2006T06, Philadelphia, Penn.: Linguistic Data Consortium.

Webber, B. L. (1988). Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th annual meeting on Association for Computational Linguistics*.

Weeds, J. (2003). *Measures and Applications of Lexical Distributional similarity*. Ph.D. thesis, University of Sussex.

Weeds, J. and Weir, D. (2005). Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, **31**(4), 439–475.

Weischedel, R., Pradhan, S., Ramshaw, L., Palmer, M., Xue, N., Marcus, M., Taylor, A., Greenberg, C., Hovy, E., Belvin, R., and Houston, A. (2008). Ontonotes release 2.0. LDC2008T04, Philadelphia, Penn.: Linguistic Data Consortium.

Wunsch, H. (2010). *Rule-based and Memory-based Pronoun Resolution for German: A Comparison and Assessment of Data Sources*. Ph.D. thesis, Seminar für Sprachwissenschaft, Universität Tübingen. `http://nbn-resolving.de/urn:nbn:de:bsz:21-opus-46044`.

Wunsch, H. and Hinrichs, E. W. (2006). Latent semantic clustering of German verbs with treebank data. In J. Hajič and J. Nivre, editors, *Proceedings of the Fifth International Workshop on Treebanks and Linguistic Theories (TLT 2006)*, pages 151–162, Prague, Czech Republic.

Yang, X. and Su, J. (2007). Coreference resolution using semantic relatedness information from automatically discovered patterns. In *ACL-2007*.

Yang, X., Zhou, G., Su, J., and Tan, C. L. (2003). Coreference resolution using competition learning approach. In *ACL-2003*.

Yang, X., Su, J., Zhou, G., and Tan, C. L. (2004). An NP-cluster based approach to coreference resolution. In *CoLing 2004*.

Yang, X., Su, J., and Yang, L. (2005a). Entity-based noun phrase coreference resolution. In *CICLING 05*.

Yang, X., Su, J., and Tan, C. L. (2005b). A twin-candidate model of coreference resolution with non-anaphor identification capability. In *Second International Joint Conference on Natural Language Processing (IJCNLP 2005)*.

Zeevat, H. (1992). Presupposition and accomodation in update semantics. *Journal of Semantics*, **9**, 379–412.

Zesch, T., Müller, C., and Gurevych, I. (2008). Using Wiktionary for computing semantic relatedness. In *AAAI 2008*.

Zielinski, A. and Simon, C. (2008). Morphisto: An open-source morphological analyzer for German. In *Finite-State Methods and Natural Language Processing (FSMNLP 2008)*.

Zinsmeister, H. (2008). Improving syntactic analysis by parse reranking. In *(Pre-) Proceedings of International Conference on Linguistic Evidence 2008*.