Fully Automatic Resolution of *It*, *This* and *That* in Unrestricted Multi-Party Dialog

von

Mark-Christoph Müller

Philosophische Dissertation
angenommen von der Neuphilologischen Fakultät
der Universität Tübingen
am 12. Juni 2008
Tübingen
2008

Gedruckt mit Genehmigung der Neuphilologischen Fakultät
der Universität Tübingen

Hauptberichterstatter:	Prof. Dr. Erhard Hinrichs
Mitberichterstatter:	Prof. Dr. Massimo Poesio
Dekan:	Prof. Dr. Joachim Knape

Für Birgit und Jette.
Ohne Euch wäre diese Arbeit
weder möglich noch sinnvoll gewesen.

**People do not remember the spoken language exactly
and so they cannot refer back to it in quite the simple way
that they can with the written language.**
Sinclair (2004, p.13)

Contents

List of Tables

List of Figures

List of Algorithms

1 Introduction

1.1 Task and Motivation

This thesis is about the automatic resolution of the pronouns *it*, *this*, and *that* in unrestricted multi-party dialog. The final part of this thesis (Chapter 7) describes experiments with an implemented system that is capable of performing this task. The system processes manually created dialog transcripts from the ICSI Meeting Corpus (Janin et al., 2003). The following Example 1 is a short fragment from one of these transcripts. The original ICSI Meeting Corpus transcript contains a semi-automatically generated segmentation which is reproduced in the example. The letters FN in the speaker tag mean that the speaker is a female non-native speaker of English. Bns003 is the identifier of the dialog that the example is drawn from.[1] The brackets and subscript numbers are not part of the original transcript.

FN083: Maybe you can also read through the - all the text which is on the web pages cuz I'd like to change the text a bit
FN083: cuz
FN083: sometimes [it]$_1$'s too long, sometimes [it]$_2$'s too short,
FN083: *inbreath*
FN083: maybe the English is not that good,
FN083: so
FN083: *inbreath*
FN083: um,
FN083: but anyways -
FN083: So I tried to do [this]$_3$ today
FN083: and if you could do [it]$_4$ afterwards [it]$_5$ would be really nice cuz I'm quite sure that I can't find every,
FN083: like, orthographic
MN021: Uh, there - there - there are couple of comments I - I have about the web pages.
FN083: mistake in [it]$_6$ or something. (Bns003)

Example 1: ICSI Meeting Corpus fragment.

For each of the six 3rd-person pronouns in Example 1, the task is to automatically identify the entity (if any) to which the speaker makes reference, and to link the pronoun to one of this entity's earlier mentions, i.e. to an antecedent. For humans, this task is often easy: it$_1$, it$_2$, and it$_6$ are used to refer to the text on the web pages, while it$_4$ is used

[1] The five dialogs from which all examples are drawn are Bed017, Bmr001, Bns003, Bro004, and Bro005, cf. Chapter 3.1.

to refer to the activity of reading this text. Humans also have no problem determining that it$_5$ is not a referring pronoun at all. In other cases, however, resolving a pronoun is difficult even for humans: this$_3$ could be used to refer to the activity of reading the text on the web pages, or to the activity of *changing* this text. The pronoun is ambiguous because in the context in which it appears, both interpretations are possible. Ambiguous pronouns are common in spoken dialog (Poesio & Artstein, 2005b). This fact has to be taken into account when building a spoken dialog pronoun resolution system.

The pronoun resolution system described in this thesis is intended to be used as a component in an extractive meeting summarization system. This summarization system takes a complete meeting transcript as its input, and creates a summary by extracting and concatenating those sections (like e.g. segments, sentences, or other structural units) that it considers to contain information that is relevant for a summary. Then, there are (at least) two ways in which the output of the pronoun resolution component can be put to use in the summarization system (cf. also Stuckardt (2003), Steinberger et al. (2007)): for improving the extraction system's *performance* (in terms of recall), and for improving the *readability* of the summary thus produced.

One way to improve the extraction system's recall is by simply substituting pronouns with their non-pronominal expressions in the transcripts prior to extraction. This way, the transcript can be made to contain a non-pronominal expression like a noun phrase or a verb phrase, instead of a pronoun with little semantic content[2], which can improve the system recall, i.e. the number of actually relevant sections that are extracted by the summarization system. Kabadjov et al. (2005), in contrast, obtained their best extraction results by specifying for each sentence whether it contained a mention of a particular automatically created anaphoric chain. One result of Steinberger et al. (2007)'s application of a similar strategy is that even a modest resolution performance results in an improvement in summarization performance. Although all these previous approaches have worked with written text and not with dialog, there is no reason why their results should not generalize to the latter.

The readability of a summary can be improved by pronoun resolution in the following way: If the final summary contains a section with a pronoun, it does not necessarily also have to contain the section with the pronoun's antecedent. If the pronoun occurs at the beginning of the summary, the pronoun is a *dangling anaphor* (Paice & Jones, 1993), i.e. it

[2] *It* is not entirely void of semantics, since it at least contrasts with the person-denoting pronouns *he* and *she*.

cannot be interpreted because it does not have an antecedent in the summary. If the section with the pronoun appears at some other position within the summary, the pronoun might be wrongly interpreted as relating to something in the preceeding context *within the summary*. These potential problems can be avoided if context-dependent pronouns in the summary are substituted with non-pronominal expressions.

1.2 Project Requirements and Choice of Corpus

Our pronoun resolution system is intended to be practically usable in a real-world setting. In this respect, it is considerably different from other implemented systems that deal with the resolution of pronouns in spoken dialog (e.g. Strube & Müller (2003), Tetreault & Allen (2004a), Tetreault & Allen (2004b)). These systems often depend on richly annotated corpora (e.g. syntactic treebanks) as input, or on manual preselection of certain pronouns that are known to be unresolvable. Their focus is not on the practical application of pronoun resolution as a preprocessing step to some other practical task. Rather, these systems are mainly used to evaluate theories about the usefulness of certain types of information. The performance of these systems tends to be reasonably good, even if it is not in the same range as that of similar systems applied to written text. In comparison, the performance of our system is bound to be much worse. This is mainly due to the unavailability of rich manual annotations and the noise in the data which results from that fact that the entire preprocessing is done fully automatically. However, we accept loss in performance in favor of practical applicability to unrestricted, unannotated dialogs. It is one of the central assumptions of this thesis that making this concession is justified because our system might turn out to have a positive effect on extractive summarization even with a comparably low absolute level of performance.

The work described in this thesis is based on manually created transcripts from a multi-party dialog corpus (see Chapter 3. These transcripts are of very high quality and fairly rich. They contain correct orthographic renderings (incl. capitalization) of the spoken utterances, punctuation, and a simple form of disfluency marking (cf. the dashes in Example 1). For the purpose of this thesis, we will make use of all of these pieces of information, even though this is at odds with our intention of having a practically usable system. For real practical applicability, an automatic speech recognition system would have to be employed, which would produce output of a lower quality than that found in the manual transcriptions. We use the rich manual transcription (thus assuming almost *perfect* accuracy of automatic speech recognition and automatic punctuation) because the current state of automatic speech recognition, especially in multi-party settings, is not sufficient yet (cf. the results in Stolcke et al. (2004)). We think that developing a system on the basis of input that is faulty even at the most basic level hampers the system's design and evaluation. This is because incorrect speech recognizer output introduces noise into our system which makes it impossible to deterministically evaluate the effects of system design decisions (incl. preprocessing). What is more, automatic speech recognition systems are getting better all the time. Automatic detection of disfluency-related information like that contained in the manual transcripts (e.g. interruption points) are fields of active research (e.g. Yeh & Wu (2006)). The same is true for automatic punctuation generation (Christensen et al., 2001). Thus, it is not entirely unrealistic to assume the type of input that we use, since there is a convergence in related fields of research to produce this type of data automatically. On the other hand, the use of an actual speech recognition system also has advantages, because it yields information about prosody and word timing that is much more detailed than that in the transcribed ICSI Meeting Corpus.

1.3 A Note on Terminology

This thesis deals with a computer system for automatically identifying the antecedents for a subset of English pronouns in spoken dialog. The design of the system is directed towards performance, i.e.: It should be able to correctly resolve as many pronouns as possible. No claims are made as to the psycholinguistic plausibility of the applied method or the linguistic correctness of the underlying model.

In the following, we use the term *referring expression* for nouns, proper nouns, and pronouns. The entity that a referring expression is used to refer to is the expression's *referent*. The relation between a referring expression and its referent is also called *mention*[3]: A referring expression *mentions* its referent, and in doing so, it functions as one of potentially several *mentions* of this referent. This definition of *mention* is more inclusive than that used in the context of the ACE project (Linguistic Data Consortium, 2005). The ACE definition of *mention* covers only expressions referring to particular types of entities (e.g. persons, geo-policitical entities like countries, or weapons). If a referring

[3] The word *mention* is used in this thesis as defined in the context of the ACE project. It is not to be confused with the same word in the opposition *mention* vs. *use*.

expression mentions a referent that has already been introduced into the discourse, this is called a *re-mention*.

New referents are introduced into the discourse by virtue of being mentioned with a referring expression for the first time. For each referent thus introduced, an entry is created in a *discourse model*. We assume this model to be much simpler than that of e.g. Webber (1991). It only contains representations of the referents and their associated mentions, but no representation of the properties and relationships that are ascribed to them in the discourse. Neither does our version of a discourse model reflect the structure of the discourse in any non-trivial way, e.g. in terms of rhetorical or intentional structure, because the analysis required for this is beyond the capabilities of current NLP systems.

1.4 Overview of the Thesis

This thesis is structured as follows: In Chapter 2 we begin by taking a closer look at the pronouns *it*, *this*, and *that*, which are the main subjects of this thesis. The chapter provides both a quantitative, corpus-based and a qualitative, functional evaluation of the characteristics of these three pronouns in spoken dialog. This motivates the choice of these particular pronouns (and the exclusion of other pronouns) for this thesis, and it helps to define the referential phenomena that we are mainly interested in. The chapter ends with a literature review about an important and well-explored functional opposition between the personal pronoun *it* on the one hand and the demonstrative pronouns *this* and *that* on the other.

Chapter 3 provides a detailed description of the ICSI Meeting Corpus, which is the multi-party dialog corpus on which our pronoun resolution system is to run. The chapter focuses mainly on two aspects: A description of the manual transcription of the corpus, which is much richer and much more accurate than what could have been acquired by the application of today's automatic speech recognition technology, and a comparison to other spoken dialog corpora that have been used for pronoun resolution. This comparison will show the much higher complexity of the ICSI Meeting Corpus, which results in considerable difficulties for pronoun resolution. The final part of the chapter gives some technical details about the XML representation used for the corpus, and about the process by means of which the corpus was converted. The XML representation is mainly necessary for the manual annotation experiments, which are the topic of Chapter 4.

Chapter 4 describes the first larger empirical part of our work. In the chapter, we apply the functional descriptions defined earlier in Chapter 2 and define annotation schemes for two extensive manual annotation experiments: The classification of referential vs. non-referential instances of *it*, *this*, and *that*, and the annotation of anaphoric relations between referential instances of these pronouns and their antecedents. Both annotation experiments are performed under a strict methodological regime, including detailed reliability checking. The chapter also describes in detail how the raw annotations yielded by each experiment are transformed into consistent data sets that could be used for machine learning. In particular for the second annotation experiment, this transformation includes the application of novel, majority-based methods. In the last part of Chapter 4, we perform an extensive descriptive analysis of the annotated corpus, which provides important information for subsequent phases of the work.

In Chapter 5, we give an overview of the current state of the art in pronoun resp. coreference resolution in written text and in spoken dialog. Although this thesis exclusively deals with the latter domain (spoken dialog), we include the domain of written text in our overview, because it is there that the vast majority of work on computational pronoun resolution has been done so far. In contrast, it will become apparent in Chapter 5 that for spoken dialog, an even remotely comparable body of work does not yet exist. In fact, this thesis marks the first attempt towards fully automatic resolution of pronouns in spoken dialog.

Chapter 6 gives detailed descriptions of two major practical parts of this thesis. The first part is the automatic preprocessing that we perform on the basis of the XML version of the ICSI Meeting Corpus, in order to turn it into a cleaner, richer, and more structured format. Automatic preprocessing includes standard tasks like sentence splitting and joining, parsing, chunking and chunk attaching, but also more specialized and challenging tasks like detection of non-referential *it*, forced time-alignment, and disfluency detection and removal. The second part is the modelling of anaphor-antecedent pairs as feature vectors, or more precisely, the definition and implementation of features in terms of which anaphor-antecedent pairs can be represented. The features used in this thesis are partly adopted or derived from related approaches (described in Chapter 5), and partly novel.

Chapter 7 integrates the findings and the practical work of the previous chapters into a running system for spoken multi-party dialog pronoun resolution. The chapter mainly describes an extensive set of machine-learning experiments in which we try to empirically determine the best settings for a couple of experimental parameters. The most important practical result of Chapter 7 is a properly evaluated, running pronoun resolution system with optimized parameters. Apart from that, the chapter also contains some qualitative result analysis and a description of alternative experiments with manually preprocessed data. The results of these experiments under *idealized* conditions, and the implications of these results for the feasibility of spoken multi-party dialog pronoun resolution in general, are discussed in the final part of the chapter.

The thesis ends with Chapter 8, in which we summarize our findings, draw some general conclusions, and outline possible ways in which the work that was only started with this thesis could be continued and improved.

2 Pronouns in Spoken Dialog: The Case of *It*, *This*, and *That*

In this chapter, we begin by motivating the choice of the pronouns *it*, *this*, and *that* as the main subject of this thesis, and the exclusion of other pronouns. We then introduce four categories of *it*, *this*, and *that* in spoken language. Some of these categories are well-known and established (like e.g. *pleonastic it*), others (like e.g. *discourse deixis*) will be (re-)defined in such a way that they are most appropriate for the purpose of this thesis. The definition of discourse deixis will also include a brief discussion of the problems of identifying semantic referent *types* for discourse-deictic pronouns. In the final part of this chapter, we will survey the major literature on the functional opposition of *it* vs. *this* and *that*.

2.1 Corpus Frequency

One reason for choosing the pronouns *it*, *this*, and *that* for this thesis is their high frequency in our corpus. In the entire ICSI Meeting Corpus ($1,013,842$ tokens), *it*, *this*, and *that* account for $50,596$ tokens, i.e. 4.99%. This makes *that*, *it*, and *this* the 6th-, 7th-, and 24th-most frequent token in the corpus, respectively. For comparison: All instances of the pronouns *he*, *his*, *him*, *she*, and *her* together account for only $2,820$ tokens, i.e. 0.28%. The detailed figures can be found in Table 1. The figures for *this* and *that* include all instances of the respective strings, regardless of the part of speech: The list does not distinguish cases where *this* and *that* are determiners resp. where *that* is a relative pronoun, complementizer, or some other part of speech.

For comparison, we also consulted the much larger British National Corpus[4] (BNC). The BNC is a balanced 100 million word corpus of both written and spoken English. The written section contains 90 million words of both imaginative and informative writing, the spoken section is made up of 10 million words of both conversational speech (4 million words) and task-oriented speech (6 million words). Based on the figures in Leech et al. (2001), we calculated relative frequencies and corresponding frequency ranks for a number of tokens in the BNC. This was done for the combined section (containing written and spoken data) and for the written and spoken section individually. In order for the counts to be comparable to the ones obtained from the ICSI Meeting Corpus,

[4] http://www.natcorp.ox.ac.uk/

Rank	**Token**	**Abs. Freq.**	**Rel. Freq.**
1	,	71603	7.06
2	.	70855	6.99
3	-	41227	4.07
4	the	33107	3.27
5	i	24583	2.43
6	**that**	**22220**	**2.19**
7	**it**	**21245**	**2.10**
8	and	19675	1.94
9	you	18881	1.86
10	's	16969	1.67
11	to	16320	1.61
12	a	13984	1.38
13	of	13922	1.37
14	so	13538	1.34
15	uh	13207	1.30
16	we	12894	1.27
17	is	11365	1.12
18	?	8579	0.85
19	do	8442	0.83
20	in	8296	0.82
21	um	8032	0.79
22	but	7854	0.78
23	have	7606	0.75
24	**this**	**7131**	**0.70**
25	one	6782	0.67
...	...	...	...
83	he	1835	0.18
...	...	...	...
248	she	380	0.04
311	his	266	0.03
325	him	246	0.03
...	...	...	...
685	her	93	<0.01

Table 1: Frequencies of selected tokens in the ICSI Meeting Corpus.

no distinction was made between different parts of speech. The results can be found in Table 2.

	Combined		**Written**		**Spoken**	
Token	**Rel. Freq.**	**Rank**	**Rel. Freq.**	**Rank**	**Rel. Freq.**	**Rank**
it	1.09	8	0.93	11	2.45	5
this	0.46	24	0.45	26	0.56	29
that	1.12	7	0.99	8	2.16	6
he	0.73	15	0.68	14	0.73	22
his	0.43	28	0.47	23	0.14	113
him	0.16	62	0.17	58	0.14	114
she	0.38	31	0.38	31	0.41	42
her	0.32	39	0.35	34	0.14	110

Table 2: Frequencies of selected tokens in the British National Corpus.

Comparing the figures in Tables 1 and 2 reveals a couple of interesting insights: Of the three pronouns of interest, *it* and *that* are consistently among the most frequent in both the combined and the individual sections of the BNC (ranking between 5 and 11), while *this* is consistently less frequent (ranking in the medium to high 20s). Comparing the relative frequencies of the three pronouns across the combined and the individual sections shows that the rate of *it* and *that* in the BNC is considerably higher in the spoken section (*it*: 2.45%, *that*: 2.16%) than in the combined (1.09% and 1.12%, respectively) or the written (0.93% and 0.99%, respectively) section. The relative frequencies of these two pronouns are strikingly similar to those found in the ICSI Meeting Corpus. The situation is similar, if less pronounced, for *this*: *this* also has a higher relative frequency in the spoken section of the BNC than in the other two sections, and this relative frequency is roughly in the same range as that found in the ICSI Meeting Corpus (ICSI Meeting Corpus: 0.7, BNC: 0.56). Thus, it turns out that the distribution of *it*, *this*, and *that* in the ICSI Meeting Corpus is very close to that found in a representative, balanced corpus of spoken language.

For the other five personal pronouns *he*, *his*, *him*, *she*, and *her*, the situation is different: The relative frequencies of *he*, *him* and *she* do not vary much across the three sections of the BNC. *His* and *her*, on the other hand, do not vary across the combined and the written section, but drop considerably in the spoken section. Table 3 contains the relative frequencies of the five personal pronouns in the spoken section of the BNC and the ICSI Meeting Corpus. The figures show that the relative frequencies for these words in the BNC consistently are several times higher than those in the ICSI Meeting Corpus.

	BNC Spoken	ICSI
Token	**Rel. Freq.**	**Rel. Freq.**
he	0.73	0.18
his	0.14	0.03
him	0.14	0.03
she	0.41	0.04
her	0.14	< 0.01

Table 3: Relative frequencies of selected tokens in BNC and ICSI Meeting Corpus.

We can conclude from the above that the high frequency of *it*, *this*, and *that* in the ICSI Meeting Corpus is probably not an artefact of this corpus. Since we find a similar distribution in the much larger spoken language section of the BNC, it is more likely that it is due to the fact that the ICSI Meeting Corpus is a corpus of spoken language. This further motivates the choice of these pronouns, because it makes the results of this thesis also relevant for other work in spoken language. The pronouns *he*, *his*, *him*, *she*, and *her*, however, are underrepresented in the ICSI Meeting Corpus. We hypothesize that this is mainly due to the fact that the topics in the ICSI Meeting Corpus tend to be very technical in nature (cf. Chapter 3.1). Thus, while their exclusion is justified for the purpose of this thesis, these personal pronouns remain a topic for further work in spoken dialog pronoun resolution, even though they are much less frequent than *it*, *this* (to some extent), and *that*.

2.2 Functional Categories

In the previous chapter, we have argued for the importance of *it*, *this*, and *that* in mere quantitative terms, and without paying attention to different grammatical functions. For example, we did not distinguish pronominal instances of *this* and *that* from determiners, or pronominal instances of *that* from determiners, relative pronouns, or conjunctions. We will now turn to a more detailed analysis of the different functions of the pronominal instances of *it*, *this*, and *that* in spoken dialog. From now on, instances of *this* and *that* that represent other parts of speech will be ignored because we assume them to be identifiable automatically. A part-of-speech tagger or a parser can be employed to recognize pronouns, determiners, relative pronouns and conjunctions by the different syntactic contexts in which they appear.

Three major referential functions of *it*, *this*, and *that* in spoken dialog will be described in Chapters 2.2.2 to 2.2.4. Before that, however, Chapter 2.2.1 will deal with the phenomenon of *non-referential* pronouns.

2.2.1 Non-referential Pronouns

As the name suggests, non-referential pronouns are pronouns that are *not* referring expressions. Non-referential pronouns do not constitute a mention of any referent. Non-referentiality is thus an important phenomenon in the context of pronoun resolution. Since non-referential pronouns are no mentions, they cannot be resolved, and thus they have to be identified and prevented from triggering a resolution attempt. Failure to do so can harm pronoun resolution precision, because a non-referential pronoun might be wrongly assigned to a referent.

Basically, two types of non-referentiality can be distinguished. A pronoun is non-referential if it is *discarded*, (Byron, 2001) i.e. if it is part of an incomplete or abandoned utterance. This type of non-referentiality is limited to spoken language, but can affect all types of expressions, not just pronouns. Discarded pronouns occur in utterances that are abandoned altogether, like the *it* in the following example from dialog Bed017.[5]

> **ME010**: Yeah. Yeah. No, no. There was a whole co- There was a little contract signed. It was - Yeah. (Bed017: four/four annotators)

If the utterance contains a speech repair (Heeman & Allen, 1999), a pronoun in the *reparandum* part is also treated as discarded. The *reparandum* is that part in the speech repair that is replaced by something that follows it (the *alteration*) and therefore, it is not part of the final utterance. In the following example from dialog Bro004, *that*$_1$, *that*$_2$, *that*$_3$, and *it*$_5$ are discarded.[6]

> **ME10**: That$_1$'s - that$_2$'s - so that$_3$'s a - that$_4$'s a very good question, then - now that it$_5$ - I understand it$_6$. (Bro004: four/four annotators)

The frequency of discarded pronouns, like that of speech disfluencies in general, is strongly influenced by factors relating to the situation of speech production (Shriberg,

[5] All examples in this thesis are drawn from the results of the manual annotation performed individually by four annotators (cf. Chapters 4.1 and 4.2). Since the annotators did not always agree, we provide for each example the rate of annotator agreement. Four/four, e.g. means that all annotators used the same tag for a given example, while three/four means that one annotator used a different one.

[6] The subscript numbers are not part of the original transcript.

1994). In many empirical works on pronouns in spoken dialog, the rate of discarded pronouns is not given, although the phenomenon is by no means rare. Schiffman (1985) reports that in her corpus of career-counseling interviews, 164 out of 838 (19.57%) instances of *it* and 80 out of 582 (13.75%) instances of *that* occur in abandoned utterances. In the corpus of task-oriented TRAINS dialogs described in Byron (2004), the rate of discarded pronouns is 7 out of 57 (12.3%) for *it* and 7 out of 100 (7.0%) for *that*.

The second important class of non-referential pronouns, which is not limited to spoken language, is *pleonastic* (or *expletive*) *it*. In its most frequent form, it occurs in an extraposition construction as a placeholder for an infinitive phrase or a sentence complement (Postal & Pullum, 1988; Kaltenböck, 2005).

> **ME013**: I think it will be interesting to do other things that aren't dumb. (Bmr001: three/four annotators)

Pleonastic *it* also takes the form of so-called *prop-it* (Quirk et al., 1991). Here, *it* is semantically empty and just fills a syntactically required position.

> **FE004**: So it seems like a lot of - some of the issues are the same. (Bed017: three/four annotators)

Pleonastic *it* is also the only type of non-referential pronoun in written text. The following figures show that it actually is a rather frequent phenomenon in written text. Evans (2001) reports that his corpus of approx. 370,000 words from the SUSANNE corpus and the BNC (written section) contains 3,170 examples of *it*, approx. 29% of which are pleonastic. Dimitrov et al. (2002) work on the ACE corpus and give the following figures: the newspaper part of the corpus (ca. 61,000 words) contains 381 instances of *it*, with 20.7% being pleonastic, and the news wire part (ca. 66,000 words) contains 425 instances of *it*, 16.5% of which are pleonastic. Boyd et al. (2005) use a 350,000 word corpus from a variety of written genres. They count 2,337 instances of *it*, 646 of which (28%) are pleonastic. Finally, Clemente et al. (2004) report that in their corpus of biomedical abstracts, nearly 44% of *it* are pleonastic. This high rate is partly explained by the fact that the text type 'scientific abstract' tends to contain stereotypical phrases which often include *it*-extrapositions or *prop-it*. For comparison, the rate of pleonastic *it* in the corpus described in Byron (2004) (task-oriented TRAINS dialogs) is 32 out of 93 (34.4%).

2.2.2 Individual Anaphoric Reference

The first category of referential instances of *it*, *this*, and *that* contains cases of *individual anaphoric reference*. These are cases where a pronoun is used to remention a referent that was introduced into the discourse by means of a noun phrase (incl. proper names). Byron (2004) calls these pronouns *NPC (noun phrase coreferential) pronouns*, while Navarretta (2004) refers to them as *IPA (individual pronominal anaphors)*. Referents of these types of pronouns will be called *NP referents*. The initial mention of an NP referent (the pronoun's *sponsor* (LuperFoy, 1991)), and all re-mentions that occurred before the current anaphoric pronoun are that pronoun's *NP antecedents*. In written text, individual anaphoric reference is the prototypical function of referential *it*, *this*, and *that*, while discourse-deictic and vague reference (cf. below) only play a marginal role. In spoken dialog, the situation is different: In (Byron, 2004), only 56% of pronouns have an NP antecedent. In (Schiffman, 1985), the rate is a comparable 54%. The following examples from the ICSI Meeting Corpus are for illustrative purposes only.

ME025: There's a sound - there's actually [sound output]$_i$ built into this thing. *Pause* But the driver - I haven't - the driver may not support [it]$_i$ yet. (Bmr001: three/four annotators)

In this example, the pronoun *it* is used to refer to the NP *sound output* introduced previously by the same speaker. Although this case seems reasonably clear, one annotator failed to annotate it in this way.

ME013: OK, what do we do with [the stuff]$_i$ on top?
ME011: Fill [it]$_i$ out. (Bmr001: four/four annotators)

In this example, all four annotators agreed on the interpretation of *it* as referring to the NP *the stuff*. Here, the coreference relation spans utterances by two distinct speakers.

ME003: Eva's got [a laptop]$_i$, she's trying to show [it]$_i$ off. (Bed017: two/four annotators)

This example is interesting because only two annotators identified the relation given above (which is assumed by the author to be the correct one), while the other two annotators interpreted the pronoun *it* as discourse-deictic (cf. below), i.e. as referring to the *fact* that Eva has a laptop.

ME013: *Pause* How are we doing on [the]
ME013: *Pause* [resources]$_i$? Disk, and -
MN007: I think we're alright, um, *Pause*, not much problems with [that]$_i$.
(Bro004: four/four annotators)

This last example is agreed upon by all four annotators, who all interpret the pronoun *it* as referring to the NP *the resources*. Here, again, anaphor and antecedent appear in utterances by distinct speakers.

2.2.3 Discourse Deixis

The second category of referential instances of *it*, *this*, and *that* contains cases of *discourse-deictic reference*. In contrast to the original definition of *discourse deixis* by Webber (1991), cf. below, we only subsume under this category cases where a pronoun is used to mention a referent that is associated with a verb phrase. Byron (2004) subsumes these types of pronouns under the name *non-NPC*, Navarretta (2004) speaks of *APAs (abstract pronominal anaphors)*. In analogy to NP referents (cf. above), referents of these types of pronouns will be called *VP referents*, and the associated verb phrase *VP antecedent*. VP referents are also sometimes called *clausally introduced entities* (Gundel et al., 2003).
As Webber (1991) notes, there is an important difference between (in our terminology) NP and VP referents with respect to the discourse model. In contrast to noun phrases, the mere occurrence of a verb phrase does not trigger the creation of a referent for the discourse model. Rather, it is assumed that in order for a VP referent to be created and added to the discourse model, the referent has to be explicitly mentioned by means of a referring expression. In the linguistic literature, this process is known as *referent coercion* (Dahl & Hellman, 1995).

Pronouns with VP antecedents belong to the class of expressions that are commonly called *discourse-deictic*. It is important to note, however, that the term *discourse deixis* resp. *discourse-deictic* is normally more inclusive. It normally also includes referring expressions that are used to refer to the content of entire stretches of text or dialog. In this thesis, we do not consider pronouns with these types of antecedents. The exact delimitation of free-form textual antecedents is often difficult even for humans (Artstein & Poesio, 2006). This is in contrast to NP and VP antecedents, where it is possible to specify rules which are based on syntactically resp. morphologically well-defined units like e.g. phrase *heads* (cf. Chapter 4.2.2). In Webber (1991), free-form textual antecedents often correspond to *discourse segments* which are defined with recourse to an elaborate model of discourse structure. No such structure is available in our model of discourse. Free-form textual antecedents can also be sequences of utterances. These can be referred back to in several ways, e.g. with *How can you say that?* or *I didn't hear you, can you repeat that?* The stretch of text that the speaker makes reference to by means of (in this case) *that* can be as difficult to identify and delimit as that of discourse segments. In addition, it is unclear what the resolution of this type of pronoun can contribute to extractive summarization, since the statement in which it occurs has a meta-communicative function. Pronouns with free-form textual antecedents are included in the class of *vague* pronouns, cf. Chapter 2.2.4.

VP referents belong to the class of entities that Asher (1993) calls *abstract objects*. Common abstract objects include e.g. states, events, propositions, situations, facts, beliefs, etc. Asher (1993) provides a very detailed and comprehensive, DRT-based study of these abstract entities. Ginzburg & Sag (2000), while focussing mainly on the semantics of questions in an HPSG-based framework, describe an ontology of semantic types on the basis of Situation Theory (Barwise, 1981; Barwise & Perry, 1983). The practical problem with these approaches resp. with the inventory of abstract objects that they provide is that they are far too complex to be easily operationalized in the form of e.g. an annotation scheme. Also, and even more importantly, as Byron (2004) notes with respect to the study by Asher (1993), the automatic identification and representation of these abstract entities on the required level of detail and subtlety is far beyond current NLP technology, making robust processing of abstract entities (resp. of the VP referents associated with them) infeasible. Against this background, it is interesting to see how earlier attempts towards the automatic resolution of pronouns with VP referents dealt with the problem of characterizing the semantic type of the referent.

As it stands, there is only one implemented system for the automatic resolution of discourse-deictic anaphors which does take the semantic type of VP referents into account.[7] The system described in Byron (2004) employs as one of the central semantic resources a manually built hierarchy of semantic types. The higher-level types of the hierarchy include e.g. *physical object*, *abstract object*, *situation*, and *event*, and Byron (2004) states that some of the more central ones are adapted from Asher (1993). What makes the hierarchy of Byron (2004) special is the fact that it is fully specified to contain types

[7] Cf. Chapter 5.3 for a more complete overview.

for all (concrete and abstract) entities in the system's domain, i.e. the TRAINS domain (cf. Chapter 3.2). As an example, Table 4 shows the branch of the type *Physical object*.

Physical-object				
	Moveable-object			
		Container		
			Boxcar	
			Tanker	
		Vehicle		
			Train	
			Engine	
		Attachment		
			Boxcar	
			Tanker	
		Cargo		
			Commodity	
			Solid commodity	
				Orange
				Banana
			Liquid commodity	
				Orange juice
	Fixed-object			
		Geographic-object		
			City	
			Factory	

Table 4: Semantic type system of Byron (2004). (*Physical object* branch).

As another example, Table 5 shows a fragment of the *Situation* branch.

Byron (2004) uses this hierarchy in two ways: First, it serves as a semantic representation of verb meanings to which different surface verbs can be mapped. For example, the surface verbs *Be*, *Get to*, *Get there*, *Get into*, and *Make it* are all mapped to the semantic type *Arrive*. The semantic type *Load*, on the other hand, is mapped to the surface verbs *Get*, *Go get*, *Load*, *Pick up*, and *Put*. Note that both *Arrive* and *Load* are sub types of *Action* (cf. Table 5). Second, the hierarchy provides a basis for the definition of semantic type restrictions for predicates, i.e. their admissible subjects and objects. For example, there are explicitly defined type restrictions stating that the semantic type for the subject of the predicate *true* (e.g. in *That's true.*) must be *proposition*, while that for the subject of the predicate *At location* (e.g. *That's in Corning.*) must be either *Event* or *Physical object*.

Since she works in a closed (and rather small) domain, Byron (2004) has the advantage of being able to completely model that domain in a bottom-up fashion, and still come

Situation			
	State		
		Appears-to-be	
		Be	
		Exist	
		Possess	
		WantNeed	
	Event		
		FailToMeet	
		Happen	
		MakeSense	
		TakeTime	
		Weather-event	
		Action	
			Arrive
			Attach
			...
			Depart
			Detach
			...
			Load
			...
			Transport
			..

Table 5: Semantic type system of Byron (2004). (*Situation* branch (fragment)).

up with a sufficiently rich and detailed semantic type hierarchy. This way, she can avoid most of the ontological problems that come with attempting a general definition of abstract types like states, events, and the like. In sharp contrast, the present thesis deals with a set of dialogs from a wider and in principle unrestricted domain. Therefore, the approach by Byron (2004) cannot be applied here. On the other hand, as will become clear in Chapter 4.2.2, the approach employed in this thesis is much more shallow than that of Byron (2004). Therefore, a distinction as fine-grained as that of Byron (2004) is not required. Instead, both for the manual annotation (Chapter 4) and for the resolution (Chapter 7), distinctions will only be made with respect to NP vs. VP referents.
There is, however, a semantic type distinction that we would like to introduce into this thesis. The classification by Moens & Steedman (1988) provides a simple and practical classification of situations and different event-types. According to Moens & Steedman (1988), events can be subdivided into event-types that differ with respect to two binary dimensions: +/- *consequent state* and *atomic* vs. *extended*.[8] The four event-types that result from the combination of these two binary features are given in Table 6. Note that the system is incomplete as it lacks e.g. propositions or facts.

	EVENTS		STATES
	atomic	extended	
+conseq.	**CULMINATION** recognize, spot, win the race	**CULMINATED PROCESS** build a house, eat a sandwich	understand, love, know, resemble
-conseq.	**POINT** hiccup, tap, wink	**PROCESS** run, swim, walk, play the piano	

Table 6: Event-types and states, reproduced from Moens & Steedman (1988).

While it is obviously much more restricted than that of e.g. Asher (1993), the classification system has the advantage of being more easily accessible from the perspective of linguistics. Another advantage is that Moens & Steedman (1988) establish a connection between the aspectual construction that a verb can appear in (mainly progressive vs. perfect), and the semantic type of the referent that it can yield when referred to anaphorically by a discourse-deictic pronoun. This observation can potentially be useful for automatic resolution. It is the basis for an operationalization for a resolution

[8] These distinctions are based on findings of Vendler (1967), cf. also Mourelatos (1978).

feature described in Chapter 6.2.1.

2.2.4 Vague Reference

We use the term *vague reference* for cases that involve a pronoun for which no clearly defined textual antecedent can be identified. Vagueness is not to be confused with referential *ambiguity*. The difference is that ambiguity involves a set of well-defined candidate referents from which none can be chosen with sufficient confidence.

Our definition of *vague* pronouns covers that part of discourse deixis where the pronoun's sponsor is a free-form textual antecedent, i.e. potentially a whole stretch of text or dialog. A similar definition of *vague* can be found in e.g. Biber (1992). Biber (1992), however, considers all pronouns without NP antecedents as vague, i.e. also discourse-deictic pronouns with VP antecedents. Free-form textual antecedents may or may not correspond to *discourse segments*. The following is a particularly clear example from dialog Bed017.

ME010: Um, so Robert, why don't you bring us up to date on
ME010: where we *are* with E_D_U?

```
(Approximately 50 intervening segments mostly by
                speaker MN015.)
```

MN015: And um. Was [that]$_{vague}$ enough of an update? (Bed017: three/four annotators)

In the above example, *that* in the last utterance is used to refer back to the entire contribution of speaker MN015 that was prompted by the request of speaker ME010. This example is not typical because the beginning of the stretch of dialog is explicitly signalled.

Our definition of *vague* also subsumes that of Eckert & Strube (2000), which is not based on the form of the antecedent, but on the nature of the referent. In the definition of Eckert & Strube (2000), *vague* pronouns are those that are used to refer to the topic of the current (sub-)dialog.

2.3 The Opposition of *It* vs. *This* and *That*

Having defined the major functional categories of *it*, *this*, and *that*, we will now look into the opposition between *it* on the one hand and the demonstratives *this* resp. *that* on the other. This opposition has been studied in a number of works, some of which explicitly deal with spoken language. In the following, we will briefly review some of this literature.

One of the earliest works on the opposition between *it* and *that* in spoken language is Linde (1979). Linde (1979) analyses a corpus of 72 short monologs (60-70 words each) in which subjects described the layouts of the various apartments that they have lived in. Linde (1979) presents a theory about the conditions under which speakers use either *it* or *that* in their descriptions. Central to this theory is the notion of *FOCUS OF ATTENTION*. Linde (1979) defines this notion in close relation to her discourse model, which in turn is closely coupled to the structure of the domain of discourse, i.e. apartments and their layout. In the simple tree-like discourse model of Linde (1979), tree nodes correspond to rooms of an apartment, and the focus is on the node representing the room that is currently being described. Based on this definition of focus, Linde (1979) observes the following regularities: Of 38 cases in which a pronoun is used to refer to the room currently being described (i.e. the referent that is currently in focus), 34 cases (89.47%) are realized by means of *it*. Of 25 cases in which a pronoun is used to refer to a referent *not* currently in focus (e.g. another room), 19 cases (76.0%) are realized by means of *that*. Linde (1979) concludes from this a strong preference for *it* to be used to refer to currently focussed referents, and a strong preference for *that* to be used to refer to referents *not* currently in focus. Another observation made by Linde (1979) has to do with reference to the current discourse topic, i.e. the apartment that is currently being described. Of 65 cases in her corpus in which a pronoun is used to refer to the apartment as a whole, as much as 62 cases (95.4%) are realized by means of *it*. Linde (1979) also identifies 15 cases in her corpus where reference is made to a previous ”statement taken as a statement” (Linde, 1979, p. 344). All of these are realized by means of *that*.

Schiffman (1985) analyzes the use of *it* and *that* in career-counseling interviews. Her corpus consists of four transcribed two-party dialogs with a total duration of 204 minutes and 31,798 words in total. It contains 1427 instances of *it* and *that*, including 244 in sentence fragments. Schiffman (1985) undertakes a statistical analysis of the contextual factors that condition the use of either *it* or *that*. She identifies 16 statistical variables pertaining to a pair of an anaphor and its antecedent. These are grouped into four so-called *accessibility* variables and seven so-called *thematic* variables. The first group contains variables which encode e.g. the (simplified) sentence distance, or the adjacency of anaphor and antecedent in terms of whether semantically compatible other expressions are present between both. Another variable in the first group takes the dialog situation into account by encoding whether antecedent and anaphor occur in the speech of the same or different speakers. The second group contains seven variables which encode e.g. the (simplified) grammatical function of antecedent and anaphor, and the form of transition between both. The remaining five variables relate to more general features, e.g. antecedent type (e.g. nominal, non-nominal, paragraph, etc.). Schiffman (1985) finds that of the 736 cases in which *it* or *that* have a nominal antecedent (incl. pronouns), 61.41% are realized with *it*, and 38.59% with *that*. She notes, however, that this correlation is not statistically significant. In contrast, the following correlations are significant: Of 109 cases in the corpus where either *it* or *that* is associated with a non-nominal constituent antecedent (mostly verb phrases), *it* is used in 30.28% of the cases and *that* in 69.72%. Similarly, if the antecedent is a sentence or a whole paragraph, *it* is used in only 11.11% of the cases, compared to 88.89% for *that*. Another finding relates to pronouns without any textual antecedents, of which her corpus contains 214 cases: Of these, 82.24% are *it*, and only 17.76% are *that*. While this correlation is also statistically significant, it cannot be interpreted since the 214 pronouns also include an unknown number of pleonastic *it*.

The work described in Gundel et al. (1993) differs in two points from the other studies discussed so far: It is neither limited to spoken language, nor to the opposition between *it*, *this*, and *that*. Gundel et al. (1993) study the distribution of all types of referring expressions in both written and spoken language, and formulate their findings in the form of the **Givenness Hierarchy**. In this hierarchy, the authors relate what they call the *cognitive status* of a referent to the form of a referring expression used to refer to it. The basic observation of Gundel et al. (1993) is that by using a particular referring expression (like e.g. a personal pronoun, a demonstrative pronoun, a demonstrative noun phrase, or an indefinite noun phrase), a speaker signals that he or she assumes the referent to have a particular cognitive status for the reader(s)/listener(s). Gundel et al. (1993) distinguish six cognitive statuses, which are given in Table 7.

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

Table 7: The Givenness Hierarchy, reproduced from Gundel et al. (1993).

Only the two highest cognitive statuses, *activated* and *in focus*, are of interest here, because according to the Givenness Hierarchy, they are conventionally associated with *this*/*that* and *it*, respectively. According to Gundel et al. (1993), a referent is *activated* for a reader/listener if it is "represented in current short-term memory" (Gundel et al., 1993, p. 278). A referent is *in focus* if it is *activated* and also at the current center of attention. This psychological definition of *focus* is identical to that of Linde (1979). Among the referents that Gundel et al. (1993) assume to generally be *in focus* are the referents of the major arguments (subject, direct object etc.), the topic of the preceeding utterance and "any still-relevant higher-order topics" (Gundel et al., 1993, p. 279).

The authors support their claims regarding the Givenness Hierarchy with an informal empirical evaluation based on small samples of naturally occurring spoken and written discourse from various sources and different languages. Details about the corpora are not provided. The English sample contains 655 referring expressions, 215 of which are *it*. Of these, a vast majority of 214 were found by Gundel et al. (1993) to refer to a referent *in focus*, and one was found to refer to an *activated* referent. In total, 246 referring expressions in the sample were used to refer to referents *in focus*, so the percentage of *it* among these is 87.0%. Gundel et al. (1993) make similar observations for *this* and *that*: The sample contains 33 expressions of this form, one of which was found to refer to a referent *in focus*, and as many as 32 of which were found to refer to an *activated* referent. In total, 150 referring expressions in the sample were used to refer to *activated* referents, so the percentage of *this* and *that* among these is 21.3%. For comparison: The percentage of definite noun phrases (*the* N) among expressions that are used to refer to *activated* referents is 63.3%.

In their empirical evaluation, Gundel et al. (1993) do not distinguish between (in our terminology) *individual*, *discourse-deictic*, and *vague* pronominal reference. Thus, their findings regarding the preference of *it* for referents *in focus* cannot be interpreted with respect to e.g. vague reference because the percentage of *topic* referents among the 215 referents of *it* is unknown. The same applies for their findings regarding the use of *this* and *that*.

Hegarty et al. (2001) study the factors that influence the accessibility of clausally introduced entities for reference with *it* resp. *this* or *that*. Entities of interest include states, events, propositions, and others, all of which were not (or not exhaustively) considered in Gundel et al. (1993). Thus, the work of Hegarty et al. (2001) complements that of Gundel et al. (1993) by using the same theoretical apparatus, i.e. the Givenness Hierarchy, to explain reference to these entities. The authors' point of departure is the observation that clausally introduced entities are generally accessible to subsequent reference with demonstratives, but comparatively inaccessible to reference with *it*. In the following example (Hegarty et al., 2001) (their Example 2) *that* is claimed to be naturally interpreted as referring to the act of destroying the leaf collection, whereas for *it*, this interpretation is claimed to be unavailable. Instead, Hegarty et al. (2001) argue that *it* is preferentially interpreted as referring to the leaf collection itself.

A: Max destroyed his leaf collection last night.
B: **That** was dumb.
It was dumb.

No empirical evaluation of this claim has been undertaken by either Hegarty et al. (2001) or the author of this thesis. In fact, contexts can quite easily be constructed in which the preference postulated by Hegarty et al. (2001) does not hold. Consider e.g.

A: Max destroyed his leaf collection last night, but it was dumb.

This example shows that more subtle factors (including overall semantic *coherence*) are at play here, which can override simple preferences based on the morphological form of a referring expression alone. On the other hand, however, the claim that demonstratives are preferred over pronouns for discourse-deictic reference is recurrent in the literature (Schiffman, 1985; Webber, 1991; Asher, 1993; Eckert & Strube, 2000; Byron, 2004; Poesio & Artstein, 2005b).[9]

Hegarty et al. (2001) offer the following explanation along the lines of the Givenness Hierarchy: Immediately after its introduction in the preceeding sentence, the act of destroying the leaf collection can be assumed to be *activated* for the reader/listener, and thus accessible to reference with a demonstrative. The referents of the syntactic arguments *John* and *the leaf collection*, on the other hand, have an even higher cognitive status, i.e. they are *in focus*. As such, *the leaf collection* is the preferred referent for *it*. Hegarty

[9] See also the analysis of the data used in this thesis in Chapter 4.2.5.2.

et al. (2001) then turn to an analysis of the different factors that bring an entity into focus. They come up with a couple of factors, including the following:
The **syntactic form** of the expression resp. the syntactic construction in which it occurs. According to Hegarty et al. (2001), there is a general preference for the thematic verb arguments (i.e. *nominal* constituents) to be in focus, rather than for the propositions or events associated with the entire utterance. Among these nominal constituents, noun phrases functioning as subjects or direct objects are more likely to bring an entity into focus than noun phrases in oblique positions.
The **world immanence** of the referent. According to Asher (1993), abstract objects show different degrees of abstractness resp. world immanence: events are the least abstract (and thus most world immanent) among all abstract entities, while propositions are the most abstract. Referents denoted by nominal constituents (i.e. NP referents in our terminology) generally have a higher world immanence than those denoted by clauses. Hegarty et al. (2001) propose that the degree of world immanence of a referent is correlated with its ability to be in focus. This is why events, but not e.g. states or propositions, are accessible to reference with *it*. They demonstrate this with an example (their Example 20) in which the preference of *it* to refer to the referent of the object argument of the preceeding sentence is overruled by the predicative context in which *it* appears.

a: John broke a priceless vase. **That** happened at noon.
b: John broke a priceless vase. **It** happened at noon.

The subject position of the verb *happen* is incompatible with the concrete noun *vase*. Actually, the predicate *happen* explicitly triggers the creation of a VP referent of type **event** from the preceeding sentence. Hegarty et al. (2001) argue that the accessibility of a clausally introduced entity with *it* depends on the type of entity, and support this with another example (their Example 21).

a: John broke a priceless vase. **That/this** was intolerable to the embassy.
b: John broke a priceless vase. **??It** was intolerable to the embassy.

Hegarty et al. (2001) claim that in the above example, the predicate *intolerable* precludes the interpretation of *it* as referring to the event, since ”an event is unchangeable once it has occurred, and thus cannot fail to be tolerated” (Hegarty et al., 2001, p. 174). Instead, they claim the predicate to force a *situation* reading, and take the markedness of the resulting sentence as indication that situations, which are claimed to be less world immanent, are less easily brought into focus.

2.4 Chapter Summary

This chapter dealt with a couple of fundamental issues pertaining to the importance and the different functions of *it*, *this*, and *that* in spoken dialog. First, we motivated the choice of these pronouns, and the exclusion of other pronouns like *he* or *she*, by showing that *it*, *this*, and *that* are among the most frequent words not only in our corpus, but also in the much larger, balanced BNC. Corpus statistics computed on the BNC also showed an even higher frequency in spoken than in written language.

Then, we turned to the exploration of the major functions of *it*, *this*, and *that* in spoken language. We began by dealing with those instances of *it*, *this*, and *that* that look like pronouns but do not actually constitute mentions. Apart from the function as a normal anaphoric pronoun, i.e. a remention of a referent that was introduced into the discourse by means of a noun phrase, we identified two other main functions that are more particular to spoken language: discourse deixis and vague reference.

For *discourse deixis*, we adopted a somewhat narrower definition: an anaphoric pronoun that is used to mention a VP referent, i.e. a referent associated with a verb phrase. This definition is in contrast to e.g. Webber (1991), who also treats those cases as discourse deixis where a pronoun is used to refer to an entire stretch of text resp. speech. The shallow approach of this thesis with respect to VP referent semantics allowed us to avoid the definition of a full-blown ontology of semantic VP referent types.

While our definition of discourse deixis is narrower than that in the literature, our definition of *vague* reference is wider than that of e.g. Eckert & Strube (2000), who treat as vague only those cases where a pronoun is used to mention the current discourse topic. We, in contrast, also include cases where reference to entire sections of text resp. dialog is made. The relevant criterion here is the absence of a well-defined textual antecedent. This definition of *vague* is mainly motivated by practical considerations of pronoun resolution.

Finally, our review of some of the literature on the opposition of *it* vs. *this* and *that* can be summed up as follows: *it* is more strongly associated with referents that are currently *in focus* than with currently non-focussed referents. NP referents functioning as thematic verb arguments are more easily rendered *in focus* than VP referents. This results in a preference for NP referents to be rementioned by means of *it*. The same is true for discourse topics, which are also among the referents that are assumed to be *in focus*. Likewise, *this* and *that* are associated with referents that are salient, but not *in focus*. Gundel et al. (1993) call this status *activated*. This status is commonly associated with VP referents, which are not normally *in focus* because the current focus is occupied by the nominal arguments of the associated verb phrase. Therefore, VP referents are more often mentioned by *this* resp. *that*.

In the next chapter, we will now turn to a description of our data.

3 Data

3.1 The ICSI Meeting Corpus

The ICSI Meeting Corpus (Janin et al., 2003) is a collection of 75 manually transcribed English-language group discussions of about one hour each. The number of participants in each discussion ranges from three to ten speakers, averaging six. Participants include male and female speakers. There is also a considerable number of non-native speakers of English. The proficiency of some of the non-native speakers is very poor, which sometimes results in highly disfluent or incomprehensible speech. The discussions are real, unstaged meetings on various, quite technical topics. Most of the discussions constitute regular weekly meetings, and as Janin (2002) points out, this often leads to a highly informal conversational style with many interrupts, asides, and jokes, because many of the participants know each other quite well.

The corpus was produced as part of the Meeting Recorder project, a project that aims at building a system for the comprehensive recording and processing of face-to-face meetings, including other modalities besides speech (Janin, 2002). The largest group of meetings in the corpus (29) deals with the Meeting Recorder project itself. The rest of the meetings deal with topics related to speech recognition, networking/Internet, and AI-oriented natural language processing.

The manual transcription is fairly rich, comprising the following information:

- **Words.** This includes both normal words in standard orthography (including capitalization and punctuation) and truncated words, the latter being transcribed with a hyphen after whatever was articulated. Truncated words correspond to what Heeman & Allen (1999) call *word fragments*.

- **Utterances.** Utterance boundaries are demarcated implicitly by means of manually added capitalization and punctuation. Utterances are stretches of words which begin with a capital letter and end with a punctuation mark. Completed utterances include (coordinated sequences of) syntactic sentences, but also non-sentential utterances (Fernández & Ginzburg, 2002). Abandoned utterances end with a single hyphen or a word fragment.

- **Non-word vocalizations.** This includes laughs, coughs, sneezes, sniffs, breaths, and similar sounds.

- **Nonvocalized sounds.** This includes door slams, microphone noises, and similar sounds.
- **Metacommments.** This category comprises various types of different information, most importantly free text comments relating to acoustic or discourse features of words, e.g. when a word or sequence of words was emphasized or articulated with special intonation.
- **Interruption points.** Hyphens (either in isolation or as the final character in a word fragment) also encode what Heeman & Allen (1999) call *interruption points*. In the simple case, an interruption point is the point at which an abandoned utterance (cf. above) ends. The fact that the next utterance is a new one (rather than a repair of the previous one) is signalled by its initial capitalization. In more complex cases, the interruption point occurs in a speech repair between the end of the *reparandum* and the beginning of the *alteration* (Heeman & Allen, 1999).
- **Non-lexicalised filled pauses.** Non-lexicalised filled pauses are transcribed in the conventional way as *uh*, *um*, etc.

The main part of each dialog in the ICSI Meeting Corpus is structured as a sequence of semi-automatically created segments. These segments are not intended to capture any linguistically relevant elements like e.g. turns, partly because the definition of turn is difficult in the context of multi-party dialog with a considerable amount of overlapping speech. Rather, the segments simply serve as time bins which provide temporal anchoring points for the transcribed data. Each segment is associated with a single speaker tag and a start and end time stamp. As a rule of thumb, the developers of the ICSI Meeting Corpus intended to insert a segment break in the current speaker's utterance whenever some other speaker started to talk (Janin, 2002). This roughly corresponds to the definition of turn that is common in discourse transcription, i.e. a change of speaker (Edwards, 2003). Given the considerable number of cases in which the speech of two or even more speakers overlapped (cf. below), this principle was not followed consequently, as it would have led to a high degree of fragmentation. As a result, the corpus contains both longer segments for utterances that clearly do overlap with other utterances, but at the same time there are also uninterrupted contributions by one speaker which are split into several segments. The latter are mainly the result of the application of a speech-nonspeech detector (Pfau et al., 2001) for the automatic creation of a preliminary segmentation. Segments are the smallest elements for which timing information

is available in the corpus. The lack of word-level time stamps, together with the inconsistent segmentation, makes it difficult to determine which words do actually overlap. It might even be argued that this lack is a serious shortcoming of the ICSI Meeting Corpus, which would be much more usable with an accurate word-level time stamping. The MapTask Corpus (Thompson et al., 1993) is an example of a data set that uses a *multi-stream* stand-off format to represent overlapping speech in a much cleaner way. Although it can be assumed that the size of the ICSI Meeting Corpus was prohibitive in that respect, word-level time stamping by means of a *forced alignment* (like e.g. in the AMI Meeting Corpus, cf. below) would have greatly improved its usability.[10]
Accurate estimates of the rate of overlapping speech in the corpus are not available. Shriberg et al. (2001) examine a subset of eight ICSI Meeting Corpus dialogs, five from the Meeting Recorder (MR) project and three from the Robustness (ROB) project. They find that the rate of overlapping speech in both subsets differs considerably. If all words are included, the rate is 17.0% for MR and 8.8% for ROB. If backchannels are not considered, the rate is 14.1% for MR and 5.6% for ROB. As Shriberg et al. (2001) note, the dialogs pertaining to the ROB project are different in that they have a main speaker to whom as much as 56% of all words can be attributed. Since this is highly untypical of the ICSI Meeting Corpus, it can be assumed that the rate of overlap in the corpus as a whole is more in the range of that for the MR project.

For the present work, only a sub-set of five randomly selected dialogs from different topics was used: Bed017 (natural language understanding), Bmr001 (the Meeting Recorder project itself), Bns003 (internet and networking), and Bro004 and Bro005 (signal processing and robustness for speech recognition).

3.2 Comparison to Other Spoken Dialog Corpora

Unrestricted multi-party dialog like that found in the ICSI Meeting Corpus constitutes a type of language data for which little discourse-level annotation or processing has been attempted so far. The few previous empirical works on pronouns in dialog have mainly used other corpora which are significantly different. Some of these other works will be mentioned repeatedly in the course of this thesis. Therefore, this chapter describes the major characteristics of these corpora, with particular emphasis on where and to what

[10] In Chapter 6.1.3, we describe a simple forced alignment algorithm that was employed to the original data in order to get at least a rough approximation of word-level time stamps.

extent they differ from the ICSI Meeting Corpus.

The **Switchboard** corpus (Godfrey & Holliman, 1993) is a collection of short telephone conversations among two native speakers of American English. The conversations were recorded in the early 90's at Texas Instruments in the course of a large project aimed at building a representative multispeaker database of telephone speech. The data was primarily intended to be used in the context of speech recognition research. Thus, the main requirement was that the collected speech should be spontaneous and natural, while the actual content was of minor importance. Each conversation covers one of approx. 70 topics that were specifically defined for the purpose. In particular, topics were chosen that are of general interest and that "tend to generate friendly differences of opinion or viewpoint, or invite exchanging of stories or shared experiences"[11]. Topics include e.g. "What short and long-term steps do you and the other caller think should be taken to improve the US budget?" or "Find out what kind of fishing the other caller enjoys. Do you have similar or different interests in the kind of fishing you enjoy?". At the beginning of each conversation, the participants were given a topic that they should discuss. The participants were unacquainted with each other, but each had previously indicated which of the 70 topics they were or were not interested in. Generally, the participants adhered to the suggested topic.

The entire Switchboard corpus consists of about 2400 conversations of approx. 6 minutes in length, conducted by over 500 different native speakers (approx. 55 % male and 45 % female). Each conversation was manually transcribed, with proper capitalization and punctuation being added. The transcript was also segmented into turns, with turn breaks being identified in the common way by speaker changes. For overlapping speech, a more sophisticated scheme was used initially, which turned out to be too impractical to be employed on a large scale. Therefore, a simpler scheme was used. In this scheme, the contribution of the speaker who is interrupting the other speaker starts a new turn, and the beginning and end of overlapping speech in both speakers' contributions are demarcated by special symbols. The rate of overlapping speech in the Switchboard corpus is rather low, as was to be expected in two-party telephone conversation. Shriberg et al. (2001) report that 12.0% of all words (including back channels like *uh-huh*) overlap. This rate drops to 7.8% if back channels are removed. For speech disfluencies, figures for the entire corpus are not available. Shriberg (1994) reports some figures for a subset of 60 conversations (40,515 words, 4,583 sentences). She finds that

[11] Switchboard User Manual (http://www.ldc.upenn.edu/Catalog/docs/LDC93S7-T/MANUAL.TXT)

in total, the corpus contains 2,586 disfluencies. In terms of sentences, 1,471 of the 4,583 sentences (32.1%) contain at least one disfluency. The utterances by both speakers are time-stamped on the word level, except for stretches of overlapping speech, where only the words spoken by one speaker have time-stamps.

The **AMI Meeting Corpus** (Carletta, 2006) is the only other dialog corpus of significant size which contains discourse-level annotation. Although pronoun resolution has apparently not yet been attempted for it, it deserves to be mentioned here as it is the only other *multi-party* dialog corpus. The AMI Meeting Corpus consists of 150 meetings (approx. 100 hours total) of manually transcribed dialog with annotations such as dialog acts, topic boundaries, named entities, but also gaze directions and other non-linguistic modalities such as hand gestures. Only about one third of the dialog data is real, uncontrolled dialog, while the rest consists of fictitious, staged dialog collected in a role-playing manner. Carletta (2006) points out that this latter data is particularly useful as it is easier to understand. This is because in the role-playing setting, the effect of the personal histories of the participants resp. their personal relationships is neutralized.

Further similarities between the AMI and the ICSI Meeting Corpus include that the former also features a significant number of non-native speakers.

Two corpora have been collected in the **TRAINS** project (Allen & Schubert, 1991) at the University of Rochester. The aim of the project was the development of an intelligent planning assistant that is conversationally proficient in natural language. Accordingly, the focus of the data collection was on relevant content rather than e.g. broad coverage of many speakers. Both corpora (**Trains91** and **Trains93**) comprise similar problem-solving dialogs between one person who is playing the role of an automatic planning assistant system, and another person who is using the system (but who is aware that the system is being played by a person). The second person, the *manager*, is told to use the system as an aid for solving a planning task. All planning tasks are related to a railroad freight system in the *Trains world*, a very simple domain consisting mainly of five cities which are partly connected by rail lines. Two of the cities have a banana resp. an orange warehouse, one city has an orange juice factory. In addition, there is a limited number of train engines and train cars. A typical task that the manager has to solve is ”The time is 12 midnight. You have to get one tanker car of orange juice to Avon, and a boxcar of bananas to Corning by 3 PM today.” The manager is told that he or she can request any type of information from the system by talking to it in a normal, unconstrained way. Thus, the dialogs are natural, but due to their being focussed on a single task, they are simple at the same time. Also, both manager and system collaborate on a common task, which has a positive effect on the dialog in terms of low rate of overlap.

Both corpora have been collected using native speakers of American English only. The **Trains91** corpus (Gross et al., 1993) consists of 16 dialogs with eight different native speakers, while the **Trains93** corpus (Heeman & Allen, 1995) contains as many as 98 dialogs with 34 different speakers. The dialogs vary greatly in length, averaging at about five to eight minutes. Each dialog was manually transcribed. In addition, the **Trains91** corpus was manually segmented into turns first, and then turns were further segmented into utterance units corresponding to intonational phrases. Overlap between speakers was marked by partly indenting the overlapping words in such a way that they appear parallel to the overlapping speech of the other speaker. In contrast, the **Trains93** corpus only underwent a more simple segmentation process, in which a segment break was added whenever a 'suitable break' occurred. The phenomena regarded as suitable include change of speaker, but also intonational phrase boundaries, and other minor phenomena if at least two of them occur at the same time. The only other segmentation principle was that the resulting segments should be neither too long (not more than twelve seconds) nor too short as to not split local phenomena. Heeman & Allen (1999) provide detailed information about disfluencies in the **Trains93** corpus. They state that approx. only 6.0% of words are disfluency-related, i.e. occur in a speech repair.

An overview of the major differences and similarities between the corpora discussed here and the ICSI Meeting Corpus can be found in Table 8.

3.3 Corpus Conversion and XML Representation

The ICSI Meeting Corpus is provided by the Linguistic Data Consortium in a simple XML format. Information other than the actual words are encoded by means of inline tags. Consider the fragment from dialog Bed017 in Figure 1.

While the XML data format in Figure 1 is adequate for representing the original information in the ICSI Meeting Corpus, it is not sufficient if additional information is to be added by either manual or automatic annotation. This is mainly due to the use of inline XML tags, which severely constrains the amount of annotation that can be added. What is more, inline annotation makes it virtually impossible to add annotation tags that span overlapping, i.e. not strictly embedded regions of text. Therefore, the orig-

	Switchboard	**AMI**	**Trains91/Trains93**	**ICSI Meeting Corpus**
two vs. multi-party	two-party	multi-party	two-party	multi-party
speakers	native English	mixed	native English	mixed
duration	ca. 6 min.	ca. 30-40 min.	ca. 5-8 min.	ca. 60 min.
medium	telephone	face to face	speech only	face to face
sentences	yes	yes	no	yes
timing level	word	word	word	segment
word fragments	yes	yes	yes	yes
interruption points	yes	yes	no	yes
purpose	conversational	(partly) unconstrained	task-oriented problem-solving	unconstrained

Table 8: Major features of some dialog corpora.

```
<Segment StartTime="1575.430" EndTime="1583.594" Participant="mn059">
 Yeah, so. So I'm - I'm not - I'm not building an <Emphasis> expert
 </Emphasis> - Uh, I want to build a smart <Emphasis> librarian, </Emphasis>
 basically that can point you to the right reference. I don't wanna compute
 the answer,
</Segment>
<Segment StartTime="1575.497" EndTime="1577.442" Participant="fe004">
 Documents that have the answers. <Comment Description="completing [mn015]'s
 utterance"/>
</Segment>
<Segment StartTime="1578.113" EndTime="1579.166" Participant="fe004">
 Mm-hmm. Right.
</Segment>
<Segment StartTime="1580.407" EndTime="1581.198" Participant="fe004">
 Right.
</Segment>
<Segment StartTime="1583.594" EndTime="1585.450" Participant="mn059">
 so it's a little bit easier for me.
</Segment>
<Segment StartTime="1584.600" EndTime="1587.560" Participant="mn015">
 <VocalSound Description="outbreath"/> <Emphasis> Well. </Emphasis> Uh, you
 have to s- still m-
</Segment>
<Segment StartTime="1587.560" EndTime="1591.643" Participant="mn015">
 understand what the content says about <Emphasis> itself, </Emphasis> and
 then match it to
</Segment>
<Segment StartTime="1590.530" EndTime="1591.439" Participant="fe004">
 Mm-hmm.
</Segment>
<Segment StartTime="1591.643" EndTime="1594.223" Participant="mn015">
 what you think the informational needs -
</Segment>
<Segment StartTime="1592.081" EndTime="1592.989" Participant="mn059">
 Mm-hmm.
</Segment>
```

Figure 1: ICSI Meeting Corpus in original XML format (fragment).

inal corpus had to be converted into a more flexible and extensible format. Since the summarization project required extensive manual annotation, we chose the data format used by the annotation tool MMAX2 (Müller & Strube, 2006). One advantage of MMAX2 in terms of data representation is that it uses stand-off annotation (Thompson & McKelvie, 1997) to associate information (in the form of so-called *markables*) with sequences of underlying text. In stand-off annotation, underlying text and markup are kept separate on the file level. This allows the representation of arbitrarily many levels of (potentially overlapping) markables on the same base data. MMAX2 also supports the representation of discontinuous markables.

In the conversion, the transcribed words and punctuation had to be separated from the inline XML tags, while the relation between both had to be preserved. In the first conversion phase, words and punctuation were tokenized, and one XML `word` element was created for each individual token. These `word` elements serve as containers for the actual word string, and they also associate a unique ID with each token. For empty tags in the input format, i.e. XML tags that do not span any tokens (like `Comment` or `VocalSound`, cf. Figure 1), special `word` elements were created. The character strings contained in these elements were derived from the values of the respective tags' `Description` attribute. The special `word` elements, which were marked by means of the attribute `meta=true` as not belonging to the actual transcribed data, are required because in MMAX2, every markable needs to be anchored in (i.e. refer to) at least one `word` element. Figure 2 shows the result of the tokenization of the XML fragment in Figure 1.

In the second conversion phase, then, all inline XML tags were processed. For each tag, a `markable` XML element was created which associated the information encoded in the tag with one or more `word` elements by means of its `span` attribute. All markables representing segments were stored on a dedicated `segment` level (Figure 3), while all other markables were stored on a generic `meta` level (Figure 4).

3.4 Annotation Environment

The manual annotations described in Chapters 4.1 and 4.2 below have been performed with the annotation tool MMAX2. MMAX2 offers a customizable display and advanced methods for rendering text, markables, and relations between markables. For both annotation tasks, a display style was used that mimicked the structure of the original transcription (Figure 1), while at the same time hiding the XML markup in favor of more

```
<word id="word_5695">Yeah</word>
<word id="word_5696">,</word>
<word id="word_5697">so</word>
<word id="word_5698">.</word>
<word id="word_5699">So</word>
<word id="word_5700">I</word>
<word id="word_5701">’m</word>
<word id="word_5702">-</word>
<word id="word_5703">I</word>
<word id="word_5704">’m</word>
<word id="word_5705">not</word>
<word id="word_5706">-</word>
<word id="word_5707">I</word>
<word id="word_5708">’m</word>
<word id="word_5709">not</word>
<word id="word_5710">building</word>
<word id="word_5711">an</word>
<word id="word_5712">expert</word>
<word id="word_5713">-</word>
<word id="word_5714">Uh</word>
<word id="word_5715">,</word>
<word id="word_5716">I</word>
<word id="word_5717">want</word>
<word id="word_5718">to</word>
<word id="word_5719">build</word>
<word id="word_5720">a</word>
<word id="word_5721">smart</word>
<word id="word_5722">librarian</word>
<word id="word_5723">,</word>
<word id="word_5724">basically</word>
<word id="word_5725">that</word>
<word id="word_5726">can</word>
<word id="word_5727">point</word>
<word id="word_5728">you</word>
<word id="word_5729">to</word>
<word id="word_5730">the</word>
<word id="word_5731">right</word>
<word id="word_5732">reference</word>
<word id="word_5733">.</word>
<word id="word_5734">I</word>
<word id="word_5735">do</word>
<word id="word_5736">n’t</word>
<word id="word_5737">wanna</word>
<word id="word_5738">compute</word>
<word id="word_5739">the</word>
<word id="word_5740">answer</word>
<word id="word_5741">,</word>
<word id="word_5742">Documents</word>
<word id="word_5743">that</word>
<word id="word_5744">have</word>
<word id="word_5745">the</word>
<word id="word_5746">answers</word>
<word id="word_5747">.</word>
<word meta="true" id="word_5748">completing
  [mn015]’s utterance</word>
<word id="word_5749">Mm-hmm</word>
<word id="word_5750">.</word>
<word id="word_5751">Right</word>
<word id="word_5752">.</word>
<word id="word_5753">Right</word>
<word id="word_5754">.</word>
<word id="word_5755">so</word>
<word id="word_5756">it</word>
<word id="word_5757">’s</word>
<word id="word_5758">a</word>
<word id="word_5759">little</word>
<word id="word_5760">bit</word>
<word id="word_5761">easier</word>
<word id="word_5762">for</word>
<word id="word_5763">me</word>
<word id="word_5764">.</word>
<word meta="true" id="word_5765">outbreath</word>
<word id="word_5766">Well</word>
<word id="word_5767">.</word>
<word id="word_5768">Uh</word>
<word id="word_5769">,</word>
<word id="word_5770">you</word>
<word id="word_5771">have</word>
<word id="word_5772">to</word>
<word id="word_5773">s-</word>
<word id="word_5774">still</word>
<word id="word_5775">m-</word>
<word id="word_5776">understand</word>
<word id="word_5777">what</word>
<word id="word_5778">the</word>
<word id="word_5779">content</word>
<word id="word_5780">says</word>
<word id="word_5781">about</word>
<word id="word_5782">itself</word>
<word id="word_5783">,</word>
<word id="word_5784">and</word>
<word id="word_5785">then</word>
<word id="word_5786">match</word>
<word id="word_5787">it</word>
<word id="word_5788">to</word>
<word id="word_5789">Mm-hmm</word>
<word id="word_5790">.</word>
<word id="word_5791">what</word>
<word id="word_5792">you</word>
<word id="word_5793">think</word>
<word id="word_5794">the</word>
<word id="word_5795">informational</word>
<word id="word_5796">needs</word>
<word id="word_5797">-</word>
<word id="word_5798">Mm-hmm</word>
<word id="word_5799">.</word>
```

Figure 2: ICSI Meeting Corpus tokens after conversion to MMAX2 format (fragment).

```
...
<markable id="markable_560" span="word_5695..word_5741" starttime="1575.430" endtime="1583.594" participant="mn059"/>
<markable id="markable_561" span="word_5742..word_5748" starttime="1575.497" endtime="1577.442" participant="fe004"/>
<markable id="markable_562" span="word_5749..word_5752" starttime="1578.113" endtime="1579.166" participant="fe004"/>
<markable id="markable_563" span="word_5753..word_5754" starttime="1580.407" endtime="1581.198" participant="fe004"/>
<markable id="markable_564" span="word_5755..word_5764" starttime="1583.594" endtime="1585.450" participant="mn059"/>
<markable id="markable_565" span="word_5765..word_5775" starttime="1584.600" endtime="1587.560" participant="mn015"/>
<markable id="markable_566" span="word_5776..word_5788" starttime="1587.560" endtime="1591.643" participant="mn015"/>
<markable id="markable_567" span="word_5789..word_5790" starttime="1590.530" endtime="1591.439" participant="fe004"/>
<markable id="markable_568" span="word_5791..word_5797" starttime="1591.643" endtime="1594.223" participant="mn015"/>
<markable id="markable_569" span="word_5798..word_5799" starttime="1592.081" endtime="1592.989" participant="mn059"/>
...
```

Figure 3: ICSI Meeting Corpus segments after conversion to MMAX2 format (fragment).

```
...
<markable id="markable_378" span="word_5712" type="emphasis"/>
<markable id="markable_379" span="word_5722..word_5723" type="emphasis"/>
<markable id="markable_548" span="word_5748" type="comment" description="completing [mn015]’s utterance"/>
<markable id="markable_139" span="word_5765" type="vocalsound" description="outbreath"/>
<markable id="markable_380" span="word_5766..word_5767" type="emphasis"/>
<markable id="markable_381" span="word_5782..word_5783" type="emphasis"/>
...
```

Figure 4: ICSI Meeting Corpus meta information after conversion to MMAX2 format (fragment).

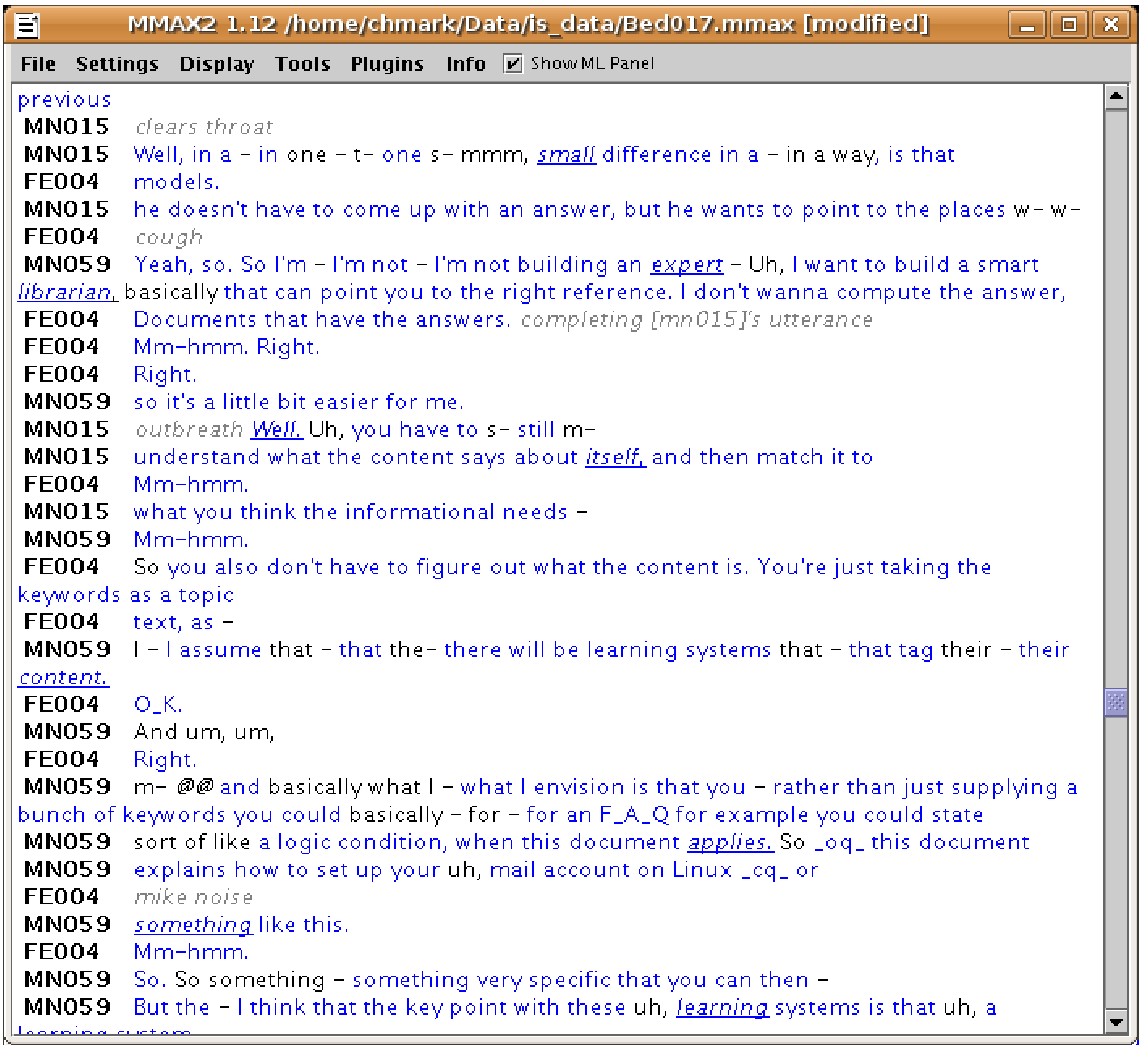
intuitive rendering styles. Figure 5 shows what the display looks like when only the converted information from the original corpus is used.

Segments are the main structural units of the display. At the beginning of each segment, the speaker tag is displayed, followed by the actual content of the segment. Note that while the readability of the text is considerably improved in comparison to the raw XML format (Figure 1), the discontinuity of overlapping segments can still make it difficult to follow the discussion. In particular, the fact that continuous utterances are apparently interrupted by other utterances that really only overlap with them (like the first utterance of FE04 or MN059 in Figure 5) is problematic. This affects what Edwards (2003, p. 325) calls ”Proximity of related events” resp. ”Time-space iconicity”. These properties describe the ability of a written transcription resp. of a display of such a transcription to adequately reflect the simultaneity or proximity of the transcribed linguistic or non-linguistic events. If e.g. the time-space iconicity is violated, *actual* adjacency of two utterances (or other events) is not reflected as *graphical* adjacency in the display.

In order to make it more easily distinguishable from the spoken text, certain types of information from the `meta` level (e.g. non-word vocalizations like *outbreath*) are rendered in light grey and italic font, while others like e.g. emphasis are only rendered in italics. The rendering of markables from the `meta` level is controlled by the display customization file in Figure 6.

The customization file contains three patterns that are applied from top to bottom to all markables on the `meta` level. The first pattern matches all markables on the `meta` level and sets their respective font style to italic. The next pattern matches meta markables of type emphasis only, and sets these markables’ font styles additionaly to underline. The last pattern matches a couple of other meta markable types and sets their display color additionally to light gray.

Information that is not currently displayed (like e.g. the start and end times of segments) is nonetheless available by selecting the respective markable with a left-click. The properties of a selected markable are displayed in a separate window of the MMAX2 tool.



previous
MN015 clears throat
MN015 Well, in a - in one - t- one s- mmm, *small* difference in a - in a way, is that
FE004 models.
MN015 he doesn't have to come up with an answer, but he wants to point to the places w- w-
FE004 cough
MN059 Yeah, so. So I'm - I'm not - I'm not building an *expert* - Uh, I want to build a smart *librarian,* basically that can point you to the right reference. I don't wanna compute the answer,
FE004 Documents that have the answers. completing [mn015]'s utterance
FE004 Mm-hmm. Right.
FE004 Right.
MN059 so it's a little bit easier for me.
MN015 outbreath *Well.* Uh, you have to s- still m-
MN015 understand what the content says about *itself,* and then match it to
FE004 Mm-hmm.
MN015 what you think the informational needs -
MN059 Mm-hmm.
FE004 So you also don't have to figure out what the content is. You're just taking the keywords as a topic
FE004 text, as -
MN059 I - I assume that - that the- there will be learning systems that - that tag their - their *content.*
FE004 O_K.
MN059 And um, um,
FE004 Right.
MN059 m- @@ and basically what I - what I envision is that you - rather than just supplying a bunch of keywords you could basically - for - for an F_A_Q for example you could state
MN059 sort of like a logic condition, when this document *applies.* So _oq_ this document
MN059 explains how to set up your uh, mail account on Linux _cq_ or
FE004 mike noise
MN059 *something* like this.
FE004 Mm-hmm.
MN059 So. So something - something very specific that you can then -
MN059 But the - I think that the key point with these uh, *learning* systems is that uh, a
learning system

Figure 5: MMAX2 annotation tool.

```
<?xml version="1.0" encoding="UTF-8"?>
<customization>
<rule pattern="{all}" style="italic=true" />
<rule pattern="type={emphasis}" style="underline=true"/>
<rule pattern="type={pause,vocalsound,nonvocalsound,comment}" style="foreground=gray"/>
</customization>
```

Figure 6: Display customization file for markables from the `meta` level.

More features of the tool will be described in Chapter 4 in the context of the two annotation experiments.

3.5 Chapter Summary

This chapter introduced the data basis used for this thesis, the ICSI Meeting Corpus. In particular, a couple of features were described which distinguish this corpus from other corpora that have previously been used in similar pronoun-related work. The ICSI Meeting Corpus consists of multi-party discussions which contain more (and more complex types of) overlap than two-party dialogs. This can only inadequately be represented in a line-based transcription. The higher number of participants also means a higher rate of speech disfluencies, because more participants mean more potential interruptions resp. interruption attempts. The rate of interruptions is also increased by the informal character of the discussions. All this, together with the entirely unconstrained and partly very technical nature of the discussions, makes the ICSI Meeting Corpus much more difficult to process than more common dialog corpora like e.g. Switchboard or Trains.

The second part of the chapter described the conversion of the original XML format to the format of the annotation tool MMAX2. The target format proved sufficiently powerful and flexible for a loss-free representation of the original data. The main reason for using MMAX2 as the target format was the fact that for the present thesis, considerable manual annotations had to be performed. These annotations are described in the next chapter.

4 Annotations

The work in this thesis encompasses two quite different annotation experiments: The first is the classification of instances of *it*, *this* and *that* into one of five classes. The main goal of this experiment is to collect data concerning the distinction between *referential* vs. *non-referential* instances of these pronouns. The second annotation experiment is based on the sub set of *referential* instances yielded by the first experiment, and deals with the identification of anaphoric relations between these pronouns and their antecedents.

For methodological reasons, all annotations were performed by four project-external, hired annotators who were naive with respect to the aim of the experiments. Two annotators (female undergrad students of computational linguistics, non-native speakers of English, henceforth annotators 1 and 2) performed both annotation experiments, while the other two (one male and one female undergrad student with no linguistics background, native speakers of English, henceforth annotators 3 and 4) were only employed in the second one. The main methodological reason for employing external annotators is that it prevents preconceived ideas from (consciously or unconsciously) interfering with the annotation. Another reason is that we wanted to check the reliability of the respective annotations, which requires at least two annotators.

4.1 Data Collection 1: Classification of *it*, *this*, and *that*

In the first of our two annotation experiments, instances of *it*, *this* and *that* were classified according to a classification scheme that was mainly based on the functional distinctions detailed in Chapter 2.2. The aim of this first experiment was twofold: First, we wanted to collect empirical data about the distribution of different types of the three pronouns in spontaneous multi-party dialog. This data can be compared to the results of similar annotation experiments on written or spoken language. One question that we were particular interested in was how reliably this type of annotation task can be performed by naive annotators. Second, we wanted to use a part of the data for building a component for the automatic detection of non-referential *it*. For pronoun resolution in written text, removing non-referential instances of *it* has become a standard preprocessing step (cf. the references in Chapter 2.2.1). For spoken dialog, this is not yet the case. Most previous work on non-referential *it* in written text, however, has not bothered to perform a reliability-controlled annotation experiment for collecting a data set of referential and non-referential instances of *it*. Instead, this decision has in most cases been made implicitly by the respective authors.

As already mentioned, the use of naive annotators ensures objectivity. In contrast, if the annotations were to be performed by the same individual who is responsible for the development of the component for the detection of non-referential *it*, this individual's decisions are likely to be influenced by ideas of how this automatic detection might work. One possible consequence of this could be that the data set tended to contain mainly clear and simple cases, leaving out harder ones.

The task of annotators 1 and 2 in the first annotation experiment was to label instances of *it*, *this*, and *that* in our five dialogs as belonging to one of the classes **normal**, **vague**, **discarded**, **extrapos it**, **prop-it**, or **other**. For reasons of clarity, we again give examples (some examples are repeated from Chapter 2).

Normal (Identifiable NP or VP antecedent)

> **ME003**: Eva's got a laptop, she's trying to show [it] off. (Bed017: four/four annotators)

> **ME013**: *Pause* How are we doing on the
> **ME013**: *Pause* resources? Disk, and -
> **MN007**: I think we're alright, um, *Pause*, not much problems with [that]. (Bro004: four/four annotators)

> **MN059**: And I think that I don't need to tell you [this]. [...] (Bed017: four/four annotators)

Vague (No identifiable antecedent)

> **ME010**: Um, so Robert, why don't you bring us up to date on
> **ME010**: where we *are* with E_D_U?

```
(Approximately 50 intervening segments mostly by
                 speaker MN015.)
```

> **MN015**: And um. Was [that] enough of an update? (Bed017: three/four annotators)

Discarded (Including utterance is abandoned)

ME010: Yeah. Yeah. No, no. There was a whole co- There was a little contract signed. [It] was - Yeah. (Bed017: four/four annotators)

ME10: [That]'s - [that]'s - so [that]'s a - [that]'s a very good question, then - now that [it] - I understand it. (Bro004: four/four annotators)

Extrapos it (Placeholder for extraposed constituent)

ME013: I think [it] will be interesting to do other things that aren't dumb. (Bmr001: three/four annotators)

Prop-it (Empty filler for required position)

FE004: So [it] seems like a lot of - some of the issues are the same. (Bed017: three/four annotators)

Other (Exophors, hedges, idiomatic uses)

FN050: And *status* would be, you know, more or less like, whether they're under construction, and - and - [or]
MN015: Uh-huh. *outbreath*
MN015: And the, uh,
FN050: [stuff like that]$_{hedge}$. (Bed017: four/four annotators)

Actually, the original tag set was more fine-grained, including categories like **exophor**, **hedge**, and **idiom** (cf. Figure 7 on page 59). However, the annotators turned out to use these categories extremely rarely only. For analysis, they are therefore conflated in the category **other**. For *this* and *that*, the annotation was to be performed on pronominal instances only, i.e. determiners, relative pronouns, adverbs, and conjunctions were to be ignored by the annotators. The categories **extrapos it** and **prop-it** are applicable to *it* only.

It is important to note that for the detection of non-referential *it* as a preprocessing step for pronoun resolution, the five-fold classification scheme (excl. the residual category **other**) could actually be simplified to a binary one. **Normal** and **vague** are subtypes of referential pronouns (incl. *it*, *this*, and *that*), while **discarded**, **extrapos it** and **prop-it** are subtypes of non-referential *it*[12]. However, we used the more fine-grained distinction because we wanted to be able to investigate the inter-annotator agreement for each of the subtypes individually. Note also that we treat **vague** *it*, *this*, and *that* as referential here even though, in the context of pronoun resolution, it would make sense to treat them as non-referential since, according to our definition in Chapter 2.2.4, they do not have an antecedent that they can be linked to. However, we follow Evans (2001) in assuming that the information that is required to classify an instance of a pronoun as a mention of e.g. the discourse topic is far beyond the local information that can reasonably be represented for the pronoun. In other words: A **vague** pronoun cannot be distinguished from a **normal** one based on mere surface features of its context. This is in contrast to **extrapos it** and **prop-it**, which appear in certain syntactic constructions only. It is also in contrast to **discarded** pronouns, which tend to be marked by the containing utterance being abandoned or disfluent.

The annotation was to be performed with the MMAX2 annotation tool. A dedicated `reference` markable level was created, which was automatically populated with markables for all instances of the strings *it*, *this*, and *that*. The two annotators received instructions including descriptions and examples for all categories, and a decision tree diagram. The diagram told them e.g. to use the *wh*-test to distinguish **extrapos it** and **prop-it** on the one hand from **normal** and **vague** on the other. The *wh*-test exploits the fact that *wh*-question formation (mostly involving *what*, cf. e.g. Eckert & Strube (2000)) is possible for referential instances of *it* only, while using a *wh*-word to ask for the 'referent' of an instance of **extrapos it** or **prop-it** causes the resulting utterance to be ungrammatical:

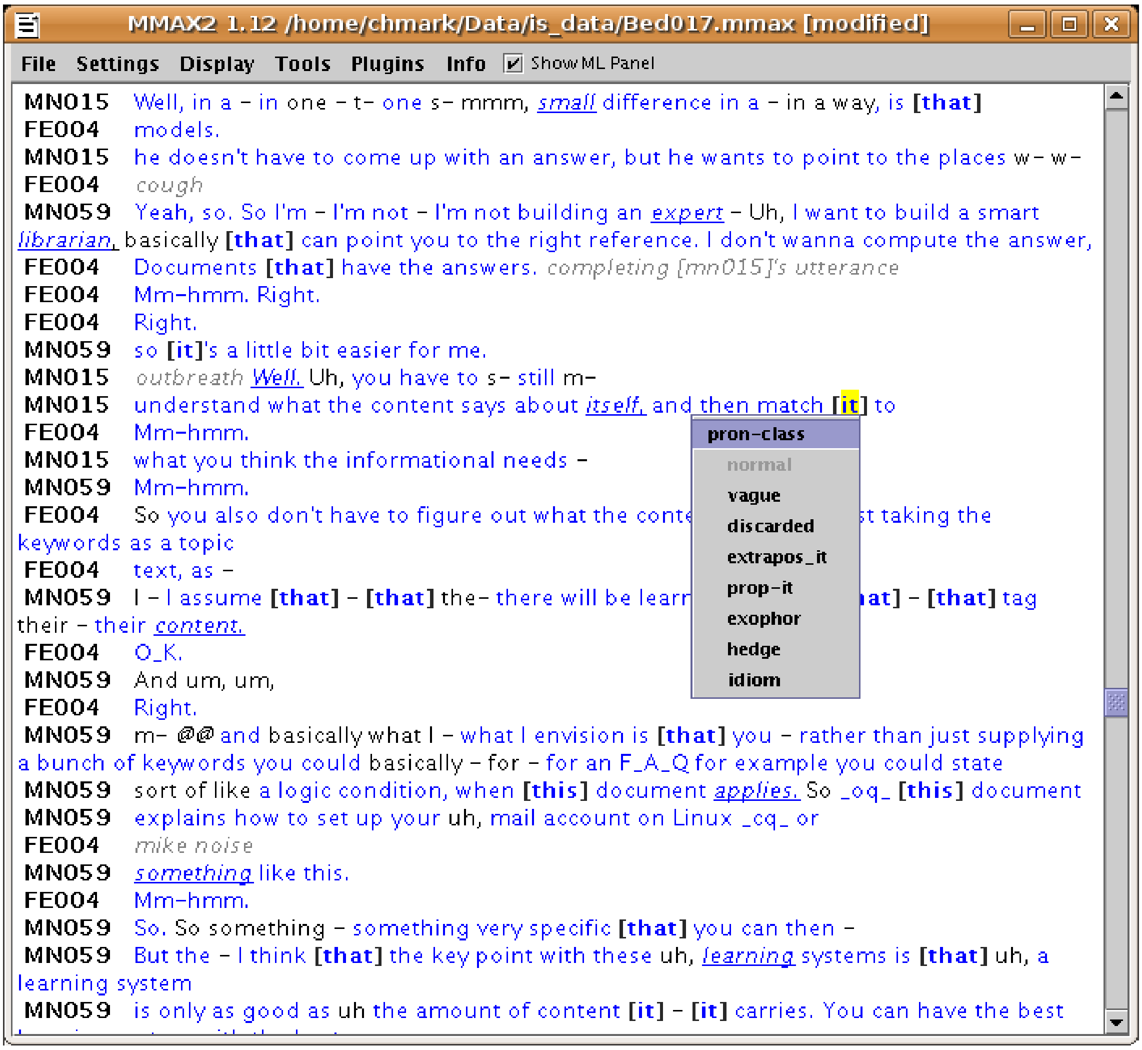
> **ME013**: *I think what will be interesting to do other things that aren't dumb.

> **FE004**: *So what seems like a lot of - some of the issues are the same.

The criterion for distinguishing between **normal** and **vague** was to use the former if an

[12] **Discarded** is the only non-referential category for *this* and *that*.

antecedent could be identified, and the latter otherwise. The annotators were also told to tag as **extrapos it** only those cases in which an extraposed element (*to*-infinitive, *ing*-form or *that*-clause with or without complementizer) was available, and to use **prop-it** otherwise. Figure 7 shows the MMAX2 display during annotation. The popup menu in the display shows the name of the MMAX2 attribute which represents the classification of the pronoun (called `pron-class`), and below this the list of possible values from which the annotator can choose. The currently selected pronoun is classified as **normal** by the annotator, which is why the corresponding value in the popup menu is disabled.



MN015 Well, in a - in one - t- one s- mmm, *small* difference in a - in a way, is [that]
FE004 models.
MN015 he doesn't have to come up with an answer, but he wants to point to the places w- w-
FE004 *cough*
MN059 Yeah, so. So I'm - I'm not - I'm not building an *expert* - Uh, I want to build a smart *librarian,* basically [that] can point you to the right reference. I don't wanna compute the answer,
FE004 Documents [that] have the answers. *completing [mn015]'s utterance*
FE004 Mm-hmm. Right.
FE004 Right.
MN059 so [it]'s a little bit easier for me.
MN015 *outbreath* *Well.* Uh, you have to s- still m-
MN015 understand what the content says about *itself,* and then match [it] to
FE004 Mm-hmm.
MN015 what you think the informational needs -
MN059 Mm-hmm.
FE004 So you also don't have to figure out what the conte st taking the keywords as a topic
FE004 text, as -
MN059 I - I assume [that] - [that] the- there will be learr at] - [that] tag their - their *content.*
FE004 O_K.
MN059 And um, um,
FE004 Right.
MN059 m- @@ and basically what I - what I envision is [that] you - rather than just supplying a bunch of keywords you could basically - for - for an F_A_Q for example you could state
MN059 sort of like a logic condition, when [this] document *applies.* So _oq_ [this] document
MN059 explains how to set up your uh, mail account on Linux _cq_ or
FE004 *mike noise*
MN059 *something* like this.
FE004 Mm-hmm.
MN059 So. So something - something very specific [that] you can then -
MN059 But the - I think [that] the key point with these uh, *learning* systems is [that] uh, a learning system
MN059 is only as good as uh the amount of content [it] - [it] carries. You can have the best

Figure 7: MMAX2: Pronoun classification task.

The two annotators 1 and 2 individually performed the annotation of the five dialogs.

The results of this initial annotation were analysed by the author of this thesis, and general problems and ambiguities in the annotation scheme were identified and corrected. The annotators then individually performed the annotation again. The results reported in the following are from this second annotation.

4.1.1 Reliability Issues

The annotation task described above has a predefined number of items (all pronominal instances of *it*, *this*, and *that*) and a small set of categories. Thus, the κ score (Carletta, 1996) is the preferred measure of inter-annotator agreement for this task. The κ score is now a de-facto standard in computational linguistics and related fields for measuring the degree of agreement between the judgments of two or more annotators. It takes into account that a certain degree of agreement might also be due to chance. κ is calculated with the following formula, where P(A) is the agreement that is actually observed between the annotators, and P(E) is the agreement between them that is expected by chance.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Normally, the value for κ ranges from 0 (only chance agreement) to 1 (perfect agreement). If the observed agreement is smaller than the expected chance agreement, the numerator of the fraction can become negative, resulting in a negative value for κ.

The κ value for a given annotation task can be interpreted as a numerical indicator of the degree of agreement. Under the assumption that all annotators worked under the same conditions, the degree of agreement can be taken as indication of how reliably the task can be performed. The rationale behind this is that if an annotation yields a κ below some threshold, this can indicate that the annotation task or the phenomenon to be annotated is inherently ill-defined or vague and thus unreliable. Carletta (1996) provides the following standard threshold values, quoted from Krippendorff (1980), which have become a quasi standard in computational linguistics. According to Krippendorff (1980), a κ value between $.67$ and $.80$ allows at most tentative conclusions, and only a value $>= .80$ can be taken to indicate actual reliability. Di Eugenio (2000) and Di Eugenio & Glass (2004) critizise the unreflected way in which these thresholds have been adopted in computational linguistics. They point out that the validity of a threshold for κ has to take the particularities of the task into account, e.g. whether the distribution of the different categories is balanced or skewed. With κ, a skewed distribution results in a very high expected agreement P(E). As will become clear in the following, in our annotation we can observe that some categories are considerably much more frequent than others, leading to a skewed distribution in our case as well.

The following Table 9 gives the inter-annotator agreement for *it*.[13]

	normal	**vague**	**discarded**	**extrapos it**	**prop-it**	**other**	**all**
Bed017	.65	.33	.94	.27	.54	.42	.62
Bmr001	.69	.21	.92	.48	.33	-.01	.63
Bns003	.59	.18	.75	.55	.21	.32	.55
Bro004	.65	-.05	.86	.75	.59	-.01	.65
Bro005	.57	-.03	.84	.58	.36	.23	.58
	.64	.11	.86	.56	.43	.20	.61

Table 9: Agreement (κ) for classification of *it* by two annotators.

In the table, the category **other** contains all cases in which one of the minor categories was selected. Each table cell contains the κ value for the respective category, calculated as described in (Fleiss, 1971). The final column contains the overall κ for the entire annotation, calculated as described in (Carletta, 1996). The final row contains the overall κ value for each category and for the entire annotation.

Table 9 clearly shows that the classification of *it* in spoken dialog appears to be by no means trivial: With one exception, κ for the category **normal** is below the $.67$ threshold. The κ for the non-referential subcategories **extrapos it** and **prop-it** is also very variable, the figures for the former being on average slightly better than those for the latter, but still mostly below the threshold. Given the fact that our annotation instructions defined what was thought to be an unambiguous indicator for **extrapos it** (i.e. the presence of an extraposed phrase), the low agreement for this class is surprising. In view of these results, it would be interesting to see similar reliability figures of annotation experiments on written texts. On the other hand, the table also shows that the detection of **discarded** instances of *it* can be done very reliably. The agreements and disagreements underlying the κ values in Table 9 can be found in Table 10 in the form of a confusion matrix.

[13] Inter-annotator agreement for *this* and *that* was not computed because we were mainly interested in the identification of different types of non-referential *it*.

	normal	**vague**	**discarded**	**extrapos it**	**prop-it**	**other**	**Anno 1**
normal	**480**	25	20	3	8	3	539
vague	33	**6**	2	2	3	2	48
discarded	13	-	**203**	1	4	3	224
extrapos it	6	-	3	**34**	22	1	66
prop-it	58	4	5	15	**56**	3	141
other	17	-	-	-	1	**4**	22
Anno 2	607	35	233	55	94	16	1040

Table 10: Confusion matrix for the classification of *it* by two annotators.

The rows in Table 10 contain the number of cases in which annotator 1 selected the category specified in the left column. Likewise, the table columns contain the respective figures for annotator 2. The bold figures along the diagonal are the number of cases in which both annotators agreed, while the numbers off the diagonal are the cases of disagreement.

The table clearly shows the huge differences in frequency among the different categories, as assigned by each individual annotator: For annotator 1, the range goes from 539 for category **normal** to only 22 for category **other**, for annotator 2, the highest and lowest frequencies are 607 and 16, also observed for **normal** and **other**, respectively.

Another advantage of visualizing the classification results in the form of a confusion matrix is that it allows to inspect which types of disagreements occur most often. It can be seen that the single most important source of disagreement in absolute terms are 58 cases in which annotator 1 classified an instance of *it* as **prop-it**, while annotator 2 classified it as **normal**. In relative terms, however, these 58 cases are only 12.8% compared to the 480 cases of agreement with respect to the category **normal**. With respect to the category **prop-it**, on the other hand, for which there are only 56 cases of agreement, this means that the number of confusions with the category **normal** is higher than the number of agreements (103.57%). It has to be noted that the confusion of these two categories is not symmetrical: The corresponding case of annotator 1 classifying an instance of *it* as **normal** and annotator 2 classifying the same instance as **prop-it** occurs only eight times, so it is likely to assume that what is underlying here is a systematic misconception of the category **prop-it** on the part of annotator 1.[14] Apart from this rather extreme case, it can be seen that quite a few of the disagreements arise from confusions of the categories **normal** and **vague** (25 resp. 33) and **extrapos it** and **prop-it** (22 resp. 15). Recall from the previous chapter that the former two categories are subtypes of referential *it*, while

[14] Note that the same appears to be true for annotator 1 and the categories subsumed under **other**.

the latter two categories are subtypes of non-referential *it*. Thus, while the distinction between e.g. **extrapos it** and **prop-it** is difficult (as can be seen from the low individual κ values in Table 9), identification of the category **non-referential** as a whole is more feasible. It follows from this that the annotation on the level of granularity that is required for the creation of a data set for the automatic detection of non-referential *it* can be done more reliably. We automatically created an alternative annotation by conflating equivalent categories, in effect giving up the more fine-grained distinctions, and calculated κ on the new annotation for comparison. The resulting agreement figures can be found in Table 11.

	normal + vague	**discarded + extrapos it + prop-it**	**other**	**all**
Bed017	.70	.79	.42	.72
Bmr001	.73	.71	-.01	.70
Bns003	.71	.73	.32	.70
Bro004	.77	.81	-.01	.77
Bro005	.62	.66	.23	.63
	.72	.74	.20	.71

Table 11: Agreement (κ) for classification of *it* by two annotators, conflated categories.

It must be stressed that category conflation with the aim of improving agreement scores on already existing annotations does not constitute a form of incorrect manipulation of the data as long as the conflated categories share a common characteristic that distinguishes them from other (single or conflated) categories. Poesio & Vieira (1998) e.g. have three naive annotators classify definite descriptions (i.e. noun phrases with the definite article *the*) in newspaper text as either **coreferent**, **bridging**, **larger situation**, or **unfamiliar**. They report a κ value of .63. They then try different ways of category conflation in order to improve the agreement. Conflating the categories **coreferent** and **bridging** on the one hand and **larger situation** and **unfamiliar** on the other yielded an improved κ of .73. When conflating the categories **bridging**, **larger situation** and **unfamiliar** into one large category and contrasting that with the category **coreferent**, κ even rose to .76.

4.1.2 Gold Standard Data Set Generation

As mentioned above, employing more than one annotator is necessary if the reliability of an annotation task is to be examined. For this, it is essential to preserve all cases of disagreement. Often, however, the collected data is to be used for more than theoretical and diagnostic purposes. This is also the case in this thesis, where the collected data is also required as training and test data. Therefore, the question arises what should be done with data instances on which the annotators did *not* agree. One option would be to ignore these instances, on the grounds that the annotators' disagreement signals that these instances are problematic or unclassifiable. However, this would reduce the number of instances in the resulting data set. In our case, e.g., even in the conflated variant, the annotators agreed in only $(480 + 25 + 33 + 6)$ (normal and vague) + $(203 + 1 + 4 + 3 + 34 + 22 + 5 + 15 + 56)$ (discarded, extrapos it, prop-it) + 4 (other) = 891 cases out of 1040. Thus, almost 15% of instances would be lost if cases of disagreement were ignored. What is more, the assumption that disagreement is a sign of problems with the classifiability is much too strong, at least for the present task. As was mentioned above, at least for some cases of disagreement there is indication that the disagreement is the result of a misconception or error on the part of one of the annotators. Therefore, ignoring data instances that involve disagreement is not a good solution.

The most common alternative is to have the annotators jointly create a gold standard version of the annotation. In this variant, cases of disagreement in the data are identified and presented to the annotators. The annotators then compare their individual decisions. In many cases, errors are obvious and can immediately be corrected. In other, controversial cases, the annotators discuss their decisions and try to reach a consensus. A data instance is ignored only in exceptional cases where no such consensus can be found.

We asked our two annotators 1 and 2 to manually create a gold standard variant of the data relating to all instances of *it* only (i.e. excluding *this* and *that*). First, a new `gold standard` markable level within the annotation tool was created which contained one markable for each instance of *it*. Each of these markables had three attributes: The category assigned by annotator 1, the category assigned by annotator 2, and the final 'gold' category. If the value in the first and second attribute was identical, it was automatically copied to the third attribute. After that, the remaining cases of disagreement could be selected within the MMAX2 annotation tool by simply querying for markables in which the third attribute was empty. The annotators then examined each of these in turn, eventually choosing the agreed-upon category or **ignore** if no consensus could be reached.

The makeup of the gold standard data set for *it* can be found in Table 12. Note the difference of 23 instances as compared to Table 10, which is mainly due to a few ignored cases. The category **other** was not considered for the gold standard data either. The figures show that a considerable number (37.5%) of *it* in our subset of the ICSI Meeting Corpus is non-referential. The gold standard data set will be used as training and test data for the development of a filter for non-referential *it*, which will be described in Chapter 6.1.1.

normal	**vague**	**discarded**	**extrapos it**	**prop-it**	**total**
588 57.8%	48 4.7%	222 21.8%	71 7.0%	88 8.7%	1017 100%
636 62.5%		381 37.5%			

Table 12: Gold standard data set for *it*.

4.2 Data Collection 2: Anaphoric Relations

The second annotation experiment focussed on the anaphoric relations between referential instances of *it*, *this*, and *that* and their textual (NP and VP) antecedents. We deliberately use the wider term *anaphoric relations* rather than *coreference*. In their criticism of the MUC-7 annotation scheme and other related schemes (cf. below), van Deemter & Kibble (2001) point out that there is no one-to-one relation between coreference proper and anaphora. Coreference is a relation between noun phrases that is established by virtue of these noun phrases having the same referent. Anaphora, on the other hand, is a relation between two noun phrases that is brought about by the second noun phrase being dependent for its interpretation on the first noun phrase. Thus, both are related, but distinct phenomena. We do not attempt coreference annotation as defined by van Deemter & Kibble (2001), because the determination of actual coreference is notoriously difficult, and, more importantly, many of the relations that we are interested in do not qualify as coreference in the strict sense. Therefore, we extend the scope of the annotation to the phenomenon of anaphora. Since we also consider antecedents that are not noun phrases (in contrast to van Deemter & Kibble (2001)), our definition is as general as that of Halliday & Hasan (1976). In the definition of Halliday & Hasan (1976), anaphora is simply a means of establishing textual cohesion by pointing back to some presupposed element in the previous discourse. In that respect, pronouns are prototypical anaphoric devices because for them, this presupposition is explicit.

In contrast to the first annotation experiment described above, the annotation of anaphoric relations is not a classification task with a small, fixed number of possible categories.

Rather, the set of options that the annotator has to choose from is equal to the set of available potential referents (including *no referent*). The size and composition of this set of potential referents, and thus the difficulty of the annotation, can vary greatly from pronoun to pronoun.

The annotation of anaphoric relations in spoken dialog is also considerably different from the annotation of the same phenomenon in written text. As will be described in the following, both tasks differ in two points in particular:

- **Ambiguity.** Traditionally, it has been a common belief in the field of anaphora in written text that it is always possible to uniquely select and annotate an antecedent for each anaphoric pronoun. Failure to do so which resulted in disagreement was ascribed to shortcomings in the annotation manual (Poesio & Artstein, 2005a). As a consequence, the infeasibility of a unique choice, i.e. referential **ambiguity**, was not an option in most annotation schemes for use with written text. For spoken dialog, however, the situation is different.
- **Non-NP antecedents.** Due to the relative rarity of discourse deixis in written text (at least in comparison to spoken dialog), annotation schemes could afford to ignore this phenomenon. In spoken dialog, as will be seen below, the situation is different, and **non-NP antecedents** have to be included in the set of potential antecedents for at least some pronouns.

In the following, we review a couple of anaphora resp. coreference annotation schemes resp. projects, paying particular attention to these two points.

4.2.1 Previous Approaches

Passonneau (1994) is one of the earliest available coreference resp. anaphora annotation manuals. It was originally developed for use with the ’Pear stories’ (Chafe, 1980), i.e. spoken narrative monologs, but Passonneau claims it to be applicable to other types of text and dialog as well. The scheme splits the annotation process into two parts. The first is the identification of discourse referential noun phrases. Passonneau provides what she calls an ”operational criterion for determining whether an NP is discourse referential” (Passonneau, 1994, p. 4). This criterion is very abstract and not easily translated into concrete instructions for annotators. The second part of the scheme covers the subsequent co-indexing of discourse referential NPs with their antecedents. Passonneau includes not only strict coreference, but also inferentially mediated relations like

Part/whole or **Causal**. Again, the processes involved in annotating these relations are very demanding of the annotator, and it seems doubtful if they can be applied reliably. All in all, the scheme strives for linguistic accuracy and completeness at the expense of simple operationalizations and thus practical applicability.

The treatment of ambiguity, however, is rather superficial, at least when compared to the ambition put into the encoding of the more sophisticated phenomena. Passonneau (1994, p. 10) states that ”[i]n most cases, a coder will have no problem assigning referential indices because the same knowledge applies to coding this feature that applies to ordinary language understanding.” She goes on to say that ambiguity is one of the cases in which this assertion does not hold, and describes how it can be encoded straightforwardly by allowing an anaphor to have more than one referential index. However, the impression remains that Passonneau believes ambiguity to be only a rare phenomenon, an opinion that might actually be correct for the ’Pear stories’. This is supported by the fact that she provides one example of pronominal ambiguity only, and even this is only temporally ambiguous because the ambiguity is resolved in the utterance directly following the one containing the ambiguous pronoun.

Passonneau (1994) treats anaphoric reference to non-NP antecedents as one of five linguistic inference relations, in this case as **Propositional inference** (Passonneau, 1994, p. 18). She also provides an example from the ’Pear stories’ of pronominal reference to a proposition, reproduced in the following (her Example 26).

6.01 [.5] But I don’t think you see the apron at first.
7.01 I don’t know if $\boxed{\text{that}}$’s important or not.

Passonneau identifies the clause *you see the apron* as the antecedent for the discourse-deictic *that*. However, no criteria are given as to how the delimitation of this clause was established. In particular, it is left open what the decision to exclude the adverbial phrase *at first* is based on, or what the antecedent would cover if the clause were more complex.

One of the best-known and most influential coreference annotation endeavors is MUC-7 (Hirschman & Chinchor, 1997). As part of a larger effort towards automatic message understanding, the MUC-7 project defined a coreference annotation task and a corresponding scheme (MUCCS) that was applicable to the MUC project’s domain of interest, i.e. business-related news or news wire reports. A major difference to the scheme by

Passonneau is that the MUC-7 scheme explicitly strives for practical applicability and efficiency, both in terms of a high inter-annotator agreement (ca. 95%) and annotation speed. Arguably, this might be one of the reasons why the scheme lacks some accuracy, cf. below. The scheme describes mainly three aspects of the annotation process: The identification of markables, the delimitation of markables, and the definition of those relations that should or should not be annotated. For the first two aspects, the scheme provides some examples and definitions, including definitions of *head of a phrase* and *maximal phrase*. All these are applicable to noun phrases only. For the relations to be annotated, the scheme mainly relies on an informal, intuitive definition, i.e. ”whether [two markables] refer to the same object, set, activity, etc.” (Hirschman & Chinchor, 1997, p. 11). Slightly more specific instructions are provided for the treatment of special phenomena like bound anaphors, apposition, and for the non-trivial distinction between coreference among types and tokens resp. functions and values. The MUC-7 annotation scheme has been much criticized (van Deemter & Kibble, 2001), and most of the criticism was aimed at the scheme's failure to properly distinguish between actual coreference and other forms of merely anaphoric relatedness.[15]

In the MUC-7 domain of journalistic, edited texts, referential ambiguity of pronouns is normally not an issue. Accordingly, the MUC-7 scheme does not contain instructions for how cases of ambiguity should be handled. However, in the context of the description of the STATUS attribute which could be used to mark optionality of coreference links, it is mentioned that this attribute could be used to allow the distinction between different causes of optionality, one of which is ”textual ambiguity” (Hirschman & Chinchor, 1997, p. 4).

The MUC-7 annotation scheme explicitly restricts itself to coreference among noun phrases. Thus, discourse deixis is not covered, though it is included as the first item in a list of possible future extensions.

The work described in Eckert & Strube (2000) also includes an empirical part covering the manual annotation of spoken dialogs from the Switchboard corpus with anaphoric relations, including discourse deixis. The description is not a manual in that it does not provide actual instructions for annotators. However, it is relevant here because it is an example of how non-NP antecedents can be handled in an annotation experiment. In addition, the annotation is the basis for the manual resolution algorithm described in

[15] All these relations were indiscriminately encoded using an `IDENT` relation.

Chapter 5.2.3.

In the context of the empirical evaluation of their algorithm, Eckert & Strube (2000) performed a series of manual annotations: dialog act segmentation, dialog act tagging, personal and demonstrative pronoun classification, and co-indexation of anaphors with their antecedents. The segmentation of turns into dialog act units was realized as a classification task in which boundaries between words were classified by Eckert and Strube according to whether or not they also constituted boundaries between dialog acts. For the annotation, dialog acts were identified with clauses. In the most simple case, a clause corresponds to a single sentence, which can be identified by its initial capitalization and sentence-final punctuation. Eckert & Strube (2000) report a very good κ value of $.92$ for the segmentation task. However, this can at least partly be ascribed to the fact that the distribution of the two categories `da-boundary` vs. `no da-boundary` is extremely skewed in favor of the latter category. In the subsequent dialog act tagging, Eckert and Strube apply a simple classification scheme on the dialog acts previously identified. In doing so, they only consider those dialog act units that they both identified independently of each other in the previous step. The annotation of anaphoric relations was split into two parts: classification and co-indexation. In the first part, Eckert and Strube classified personal and demonstrative pronouns in their corpus as **Individual**, **Discourse Deictic**, or **Vague**. For personal pronouns, an additional category **Inferrable-Evoked** was used.[16] For personal pronouns, the classification task yielded a κ value of $.81$, while κ for demonstrative pronouns was $.80$. In the second part of anaphoric annotation, Eckert and Strube co-indexed personal and demonstrative pronouns with their nominal or clausal antecedents. In this process, only those anaphors were used whose classification Eckert and Strube agreed upon in the previous step. After the annotation, Eckert and Strube created a reconciled version of the annotation and measured the accurracy (in percent agreement) of the individual annotations with the reconciled version. The values ranged from 85.7% to 98.4%.

Eckert & Strube (2000) do not provide an explicit means for the encoding of referential ambiguity. Rather, for each referential personal or demonstrative pronoun, exactly one antecedent had to be specified. While it seems odd to assume that all referential pronouns in their dialog corpus are non-ambiguous, this is actually at least somewhat justified because the set of referential pronouns that was considered for co-indexation is heavily pre-filtered. As mentioned above, it contains only those pronouns whose classi-

[16] This category covers instances of (mostly) *they* like that in Eckert and Strube's example 18: "... in **the Soviet Union**, **they** spent more money on, um, what do you call, um, military power than anything."

fication Eckert and Strube agreed upon, while all other pronouns were dropped. Thus, at least pronouns for which a potential ambiguity already lead to a disagreement in their classification (e.g. **Individual** vs. **Discourse Deictic** or **Individual** resp. **Discourse Deictic** vs. **Vague**) apparently were never even considered as anaphors for which an antecedent had to be specified.

One of the major features of the algorithm described in Eckert & Strube (2000) is the ability to distinguish between, and resolve, normal and discourse-deictic anaphors. Thus, their annotation includes provisions to specify non-NP antecedents as the antecedents of discourse-deictic pronouns. Referents of discourse-deictic pronouns are called *abstract referents* by Eckert & Strube (2000). They include propositions, facts, events, etc., but it is worth noting that the characterization of the semantic type of a given abstract referent is beyond the scope of the annotation performed by Eckert and Strube. All abstract referents have in common that they are associated with dialog acts. Thus, the annotation of antecedents for discourse-deictic pronouns is done by co-indexing the pronoun with the dialog act that was identified manually in one of the previous annotation steps. The problem of antecedent delimitation for discourse-deictic pronouns is thus shifted to the more general problem of dialog act segmentation. The description of the actual co-indexation process in Eckert & Strube (2000) is somewhat glossed over, and a couple of things are unclear. E.g., if only those dialog acts were used in later annotation steps that the annotators agreed upon in the very first step, how were those pronouns handled for which both annotators agreed in classifying them as **Discourse Deictic**, but for which no clausal antecedent was available because the delimitation of this antecedent was not agreed upon in the first annotation step?

The annotation scheme in Byron (2003) is different from the one by Passonneau, Eckert and Strube, and the MUC-7 scheme in that it was explicitly designed for the annotation of anaphoric pronouns (including discourse deixis) in spoken dialog from the Trains93 corpus by naive annotators. It is also the scheme that underlies the work described in Chapter 5.3.2.

Byron uses the term *linguistic* antecedent for the textually given antecedent (i.e. the *sponsor*), in order to distinguish it from what she calls the *semantic* antecedent, which is the actual referent of the pronoun.

An important methodological prerequisite, stated by Byron (2003), was that the annotation should be performed by naive annotators with no prior training in annotation or computational theories of reference. This is reflected in her annotation scheme in the fact that only moderately difficult decisions are required from the annotators. In spite of the claim to base the annotation on the interpretations of naive annotators, Byron herself had some influence on the outcome of the annotation, because she was one of the two annotators in the annotation of eleven of the 19 Trains93 transcripts in her corpus.[17] In the subsequent data reconciliation process, cases of disagreement were discussed by both annotators in order to find a correct answer. It can be assumed that the expertise of Byron had a positive effect on the outcome of the reconciliation and thus on the quality of the data ultimately produced. Byron herself states that a consensus could be found in all but a very few cases. In these rare cases, the value of a randomly selected annotator was chosen as the correct answer.

Byron (2003) made provisions in her scheme to allow the annotators the encoding of referential ambiguity in the **Semantic Antecedent** feature. If the annotators found more than one referent to be plausible, and if they could not decide on any single referent, they had two options: They could set the value to *ambiguous*, or specify several (not more than two or three) candidates in preference order. If the annotators felt a pronoun to be ambiguous among more than three referents, they were instructed to annotate it as *ambiguous*. However, Byron notes that none of the annotators ever specified more than one referent for a pronoun. Likewise, the value *ambiguous* also appears to have been selected extremely rarely only. The κ values for the **Semantic Antecedent** feature are $.71$ for all pronouns and $.56$ resp. $.82$ for demonstrative resp. personal pronouns only. This rather high agreement might be seen as the result (at least to some extent) of the fact that the semantic antecedent was specified by the annotators in the form of a textual description. In other words: For the **Semantic Antecedent**, the annotators were not required to identify and delimit a particular stretch of dialog. For this, a different feature was used, the **Linguistic Antecedent**. The κ values for this feature are $.66$ for all pronouns and $.37$ resp. $.77$ for demonstrative resp. personal pronouns only. It can also be assumed that the limited complexity of both the Trains93 domain and the respective dialogs considerably simplified the task of semantic antecedent identification.

Byron is also explicitly interested in pronominal reference to propositions etc. These were also annotated in the way described above, i.e. by specifying a textual representation of the referent. Byron subsumes pronominal mentions of higher-order referents like events or propositions under the category of pronouns with no explicit antecedent in

[17] The remaining eight transcripts were annotated by one annotator only. It is unclear if this was Byron herself or one of the naive annotators.

the dialog. Consequently, no **Linguistic Antecedent** has to be specified for these cases, and the scheme does not have to provide instructions for how textual antecedents for discourse-deictic pronouns have to be identified and delimited. Also, the κ values for **Linguistic Antecedent** reported above benefit from the fact that they are calculated on the basis of nominal antecedents only.

The work described in Poesio & Artstein (2005a) marks the first attempt to systematically annotate ambiguity in spoken dialog, including ambiguity of discourse-deictic pronouns. Poesio & Artstein (2005a) start from the observation that previous work in the annotation of discourse regarded ambiguity only as a source of unwanted disagreement which should (and in fact could) be eliminated. They contrast this with the view that – particularly in discourse – genuine ambiguity does obviously exist, and that methods should be developed to handle it. Poesio & Artstein (2005a) and Artstein & Poesio (2006) performed a series of annotation experiments with the MMAX2 annotation tool and with a large number of naive annotators (up to 20) on dialogs from the Trains91 corpus. Their motivation was to exchange ”the highly knowledgeable opinions (and prejudices) of experts with the collective wisdom of many speakers” (Artstein & Poesio, 2006, p. 56). In the first part of the annotation, annotators had to classify all noun phrases (including pronouns and excluding temporal noun phrases) as belonging to one of the categories **none**, **phrase**, **segment**, or **place**. All noun phrases were already identified in the data, so the annotators did not have to define or create markables for them. The first category **none** applied to non-anaphoric or non-referential noun phrases. The second and third category was to be assigned to noun phrases that were coreferent with a noun phrase antecedent (**phrase**) or with a segment (**segment**), i.e. a non-nominal, antecedent. The distinction between **phrase** and **segment** is roughly equivalent to that between **Individual** and **Discourse Deictic** anaphor (Eckert & Strube, 2000), cf. above. The fourth category (**place**) simply served to distinguish mentions of place names in the Trains91 domain, which are frequent but considered uninteresting, from other noun phrases.

In accordance with their particular interest in referential ambiguity, Poesio & Artstein (2005a) allow their annotators to specify arbitrarily many antecedents in case of explicit ambiguity, i.e. ambiguity actually perceived by the annotator.[18] Other than the scheme of Byron (2003), their scheme does not allow to express a preference ordering for the

[18] Poesio & Artstein (2005a) distinguish *explicit* ambiguity from *implicit* ambiguity which becomes evident only when interpretations of different annotators are compared and found to be distinct.

various candidates.
Antecedents are specified in the scheme of Poesio & Artstein (2005a) by letting the anaphoric noun phrase point to its antecedent(s). This is implemented using a special feature of the MMAX2 annotation tool. For anaphors of the category **phrase**, the target of this pointing relation was the prefabricated noun phrase markable. For discourse-deictic (i.e. **segment**) anaphors, Artstein & Poesio (2006) describe two distinct ways of pointing to their antecedent(s). In the less constrained variant, annotators were allowed to mark arbitrary regions of text and create markables for those, which then served as the pointing target. The high variance that was to be expected in this task was countered in the experiments by Artstein & Poesio (2006) by employing as many as 20 annotators. In the more constrained variant, the domain of antecedent markables was limited to utterances as defined in the corpus. For this task, only four annotators were employed.

4.2.2 Our Approach

Anaphoric pronouns in unrestricted multi-party dialog like the ICSI Meeting Corpus have not yet been extensively studied. Therefore, apart from acquiring test and training data for automatic pronoun resolution, our data collection also had the aim of gathering empirical data about the phenomenon.
The annotation scheme that we used for the annotation of anaphoric *it*, *this*, and *that* was mainly shaped by

1. methodological considerations,
2. requirements of the resolution task, and
3. practical feasibility and resource constraints.

The single most important methodological consideration was that we wanted to employ naive annotators only. The reason for this is the same as that for employing only naive annotators for the classification of *it*, cf. Chapter 4.1. Using only naive annotators precludes the use of a highly sophisticated annotation scheme, since that might be too demanding and presuppose a level of linguistic knowledge that cannot be expected from naive annotators with only a limited amount of training.

Ideally, the requirements of the resolution task should as little as possible influence the design of the annotation scheme that is used for producing the data. Rather, the annotation should be optimized to do justice to the annotated phenomenon. However, the resolution task as we define it has two major requirements on the format of the data. The first requirement is that for every anaphoric pronoun, *exactly one* textual representation must be available in the preceeding context of the dialog. In the terminology of Byron (2003), this means that a single **Linguistic Antecedent** must be specified for every **Semantic Antecedent**, including those of discourse-deictic pronouns. The second requirement is that these textual antecedents must be identifiable fully automatically. This requirement follows from the aim of building a practically usable system which does not depend on manually preprocessed data. Even under the concession of using all transcription information in the ICSI Meeting Corpus, this aim precludes the use of any non-trivial linguistic units (clauses, turns) as antecedents, because, as described in Chapter 3, the corpus itself contains a simple segmentation only.

Finally, practical feasibility and resource constraints limited the number of annotators that could be employed. On the one hand, access to sufficiently capable and motivated students was limited. On the other hand, given the fact that a considerable amount of data had to be annotated and that annotation was time-consuming, limited resources prohibited the employment of additional annotators for the required considerable length of time. Although it would have been preferable to employ only native speakers of English, only two native English speakers (3 and 4) could be recruited for the annotation.

As was already mentioned, the annotation was performed by four naive project-external annotators. Annotators 1 and 2 had already been employed in the first annotation task, cf. Chapter 4.1. The annotators received an annotation manual that in the first part explained and illustrated the basic notions of coreference, anaphora, and discourse deixis. These instructions were deliberately kept simple, in consideration of the fact that the annotators were no linguistics experts. The notion of *antecedent*, e.g. was defined rather informally as the expression that the pronoun referred to. This practice of using a larger number of naive – rather than only two, highly trained – annotators was inspired by and shared the rationale of the experiments in Poesio & Artstein (2005a) and Artstein & Poesio (2006). The major difference between our annotation and that of Poesio and Artstein is that our annotators were instructed to choose the single most plausible interpretation in case of perceived ambiguity. Since the annotation was limited to the pronouns *it*, *this*, and *that*, the annotation task and thus the training requirements were considerably simplified. Other than in coreference resp. anaphora annotation covering all types of noun phrases, our annotators were not required to solve tasks like determining the referentiality or anaphoricity of noun phrases (Passonneau, 1994; Hirschman & Chinchor, 1997).

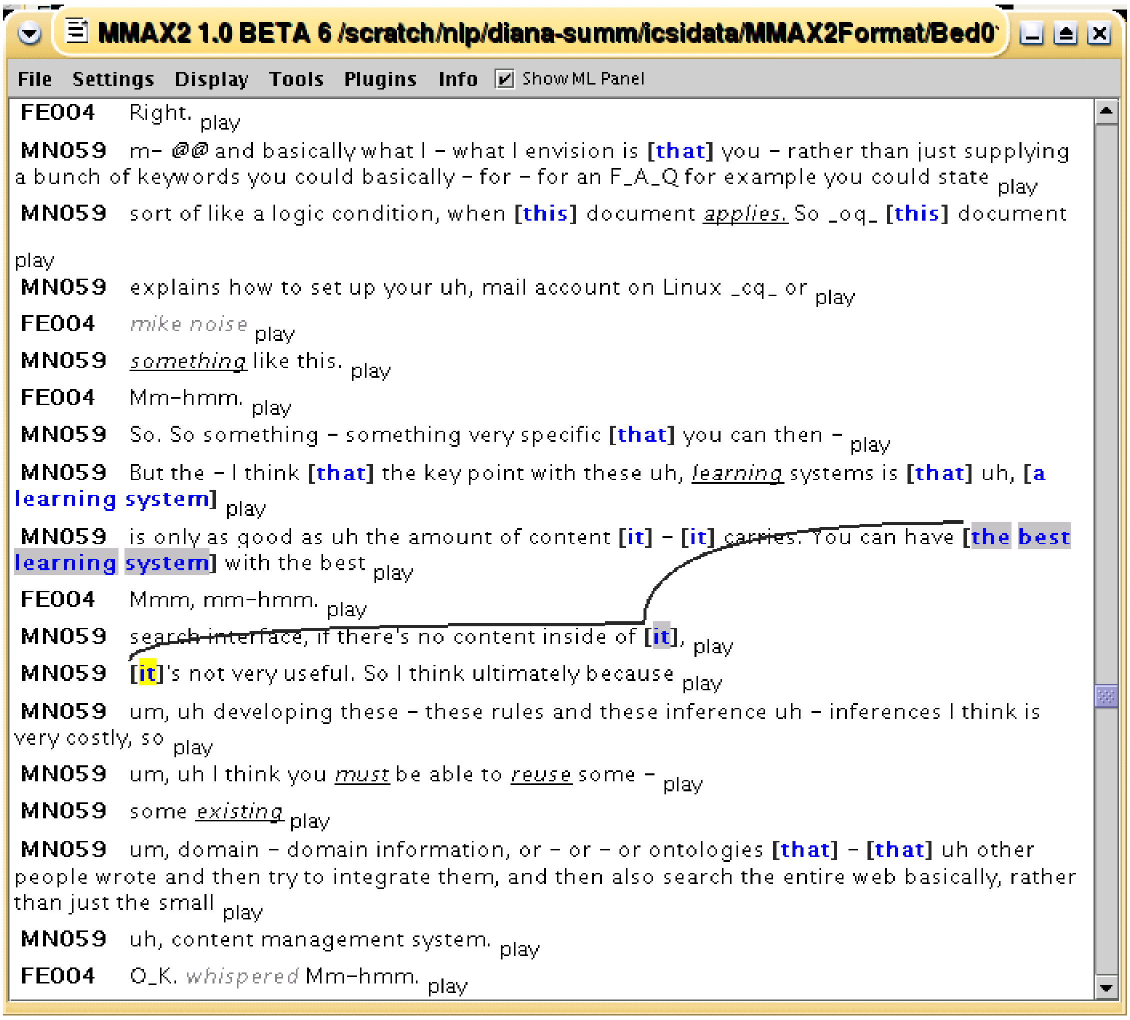
In the second part, the annotation manual described how markables were to be created (if necessary) and linked in the MMAX2 annotation tool. Linking of markables was implemented by means of a special feature of the tool. One of the markable relations supported by the MMAX2 data format is the so-called *markable set*. It can be used to model equivalence relations between two or more markables. Membership of a markable in a particular set is expressed by means of a special attribute which has the same numerical value for all markables in the same set. In addition, all markables in a markable set are ordered in discourse order[19]. The use of an equivalence relation (i.e. a transitive relation) allows us to maintain the transitivity between anaphoric markables in a set with three or more elements, while the markable set *ordering* represents the element of *direction* in the anaphor-antecedent relation. Markable sets can be converted into anaphoric chains by linking each anaphor to its immediate, i.e. most recent antecedent.

In MMAX2, the mechanisms for adding a markable to or removing a markable from a markable set are hidden from the annotator. Annotators only have to select a markable (normally the anaphor) by left-click, and then right-click on the antecedent. Figure 8 shows the MMAX2 display during annotation of anaphoric relations.

Figure 8 shows a three-element markable set. The elements of this set are linked graphically in discourse order to form an anaphoric chain. If an anaphoric chain contains a non-pronominal antecedent (here: the NP *the best learning system*), this is always the first element in the chain. One consequence of this annotation policy is that longer chains that include anaphoric definite noun phrases are not annotated as a whole. Rather, they are split into several shorter chains, each of which begins with a non-pronominal antecedent.

For all annotators, markables for all occurrences of the strings *it*, *this*, and *that* were created automatically. From among the pronominal instances, the annotators then identified normal, vague, and non-referential pronouns. For normal ones, they also linked them to their most recent antecedent, either a noun phrase (NP antecedent, including pronoun) or a verb phrase (VP antecedent). If annotators found a pronoun to be referential, but lacking an antecedent, they were instructed to use the category `vague`. In

[19] I.e. in order of appearance in the dialog transcript, which is not necessarily the correct chronological order.


MMAX2 1.0 BETA 6 /scratch/nlp/diana-summ/icsidata/MMAX2Format/Bed0
File Settings Display Tools Plugins Info ☑ Show ML Panel

FE004 Right. play

MN059 m- @@ and basically what I - what I envision is [that] you - rather than just supplying a bunch of keywords you could basically - for - for an F_A_Q for example you could state play

MN059 sort of like a logic condition, when [this] document *applies.* So _oq_ [this] document play

MN059 explains how to set up your uh, mail account on Linux _cq_ or play

FE004 *mike noise* play

MN059 *something* like this. play

FE004 Mm-hmm. play

MN059 So. So something - something very specific [that] you can then - play

MN059 But the - I think [that] the key point with these uh, *learning* systems is [that] uh, [a learning system] play

MN059 is only as good as uh the amount of content [it] - [it] carries. You can have [the best learning system] with the best play

FE004 Mmm, mm-hmm. play

MN059 search interface, if there's no content inside of [it], play

MN059 [it]'s not very useful. So I think ultimately because play

MN059 um, uh developing these - these rules and these inference uh - inferences I think is very costly, so play

MN059 um, uh I think you *must* be able to *reuse* some - play

MN059 some *existing* play

MN059 um, domain - domain information, or - or - or ontologies [that] - [that] uh other people wrote and then try to integrate them, and then also search the entire web basically, rather than just the small play

MN059 uh, content management system. play

FE004 O_K. *whispered* Mm-hmm. play

Figure 8: MMAX2: Anaphor annotation task (NP antecedent).

case of perceived ambiguity, the annotators were instructed to select the antecedent corresponding to the most plausible interpretation. If no such interpretation was available to them, they were instructed to classify the pronoun as `not_set`. All other instances of the strings *it*, *this*, and *that* remained unannotated.

From their earlier annotation, annotators 1 and 2 were already familiar with the dialogs. In particular, they had already collaborated in the creation of the gold standard data set for non-referential *it* (Chapter 4.1.2). This posed a potential problem for the annotation of anaphoric relations, because annotators 1 and 2 in their discussions might have reached a common interpretation of some potentially ambiguous pronouns that was different from their previous, individual and unbiased interpretations. On the other hand, there was a break of several weeks between the two annotation tasks, and annotators 1 and 2 had no access to their previous annotations while doing the anaphoric annotation. Still, the different level of acquaintance with the data of annotators 1 and 2 as compared to annotators 3 and 4 has to be taken into account.

Markables for non-pronominal antecedents (i.e. full noun phrases and VP antecedents) had to be created by the annotators via the MMAX2 GUI by holding the left mouse button and dragging the mouse over a portion of text. For NP antecedents, the annotators were instructed to create markables for *simple* noun phrases only, i.e. spanning the phrase head plus any premodification, but without postmodification.[20] This is the common way of creating markables for noun phrases. For VP antecedents, the annotators were instructed to create markables for the verb phrase *head* (i.e. the verb) only, and not for the entire phrase or clause/sentence. To our knowledge, this type of minimalist definition of non-NP antecedents has not been used before. It was mainly chosen in order to minimize the disagreements between annotators that are caused by differing markable delimitations (Poesio & Artstein, 2005a), and because it obviously meets the criterion of being automatically identifiable. Of course, the verb alone does not sufficiently represent an entire clause or sentence. In a dependency-based view of language, however, the verb can be seen as the central element from which all other elements are (directly or indirectly) dependent (Mel'čuk, 2003). In the context of this thesis, the determination of these elements is regarded as a task outside the scope of anaphor resolution, and it is assumed that the identification of the verb provides sufficient information for the subsequent identification of the corresponding larger unit.

[20] The manual given to the annotators avoided to use these technical terms. Instead, it made use of illustrative examples.

4.2.3 Reliability Issues

The annotation of anaphorically related expressions by grouping into equivalence classes is a task for which classification-based measures of inter-annotator agreement (like e.g. κ as described in Carletta (1996)) are not appropriate (Artstein & Poesio, 2005). This is because, as already mentioned in the beginning of Chapter 4.2, the set of candidate antecedents for every single anaphor in a dialog is not a constant, closed list. Rather, it is heavily dependent on the respective anaphor and its context, and an enumeration of *all* candidate antecedents for *all* anaphors in a dialog (as would be required for the application of κ) is infeasible. In addition, defining the notion of the *correct* antecedent for a given anaphor is difficult, because two antecedent candidates might or might not be equivalent antecedents for a given anaphor, depending on how they themselves are annotated. This is mainly due to the fact that coreference is a *transitive* relation, i.e.: If some expression A is coreferent with some expression B, and expression B is coreferent with some expression C, then expression A is also coreferent with expression C.

Another quantitative measure for comparing anaphora annotations which does take set-related properties like transitivity into account is the one described in Vilain et al. (1995). This is also one of the standard measures for coreference resolution evaluation. The measure can be used to quantitatively compare two sets of markables (regarded as *key* and *response* respectively) with each other on the basis of how many links have to be added or removed in order to make both annotations identical. Since it is intended as an evaluation measure, the measure described in Vilain et al. (1995) assumes one annotation (the *key*) to be correct. Links in the *response* are regarded as correct if they are also in the *key*, either directly or transitively. The measure yields two numerical values between $.0$ and 1.0: The *precision* (P) is the proportion of the number of correct links in the *response* to the number of all links in the *response*. The *recall* (R) is the proportion of the number of correct links in the *response* to the number of all links in the *key*, i.e. to all correct links. Precision and recall are normally combined into the single *F-measure* calculated according to the following formula.[21]

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

[21] β is a weighting factor which is normally set to 1, giving equal weight to both P and R (Jurafsky & Martin, 2000).

Precision and *recall* are symmetrical, i.e. the value of *F* is the same regardless of which of the two sets is taken as the *key* and which as the *response*. Therefore, *F-measure* calculated according to Vilain et al. (1995) could in principle be used as a quantitative measure of inter-annotator agreement on markable sets. This has been proposed, among others, by Hirschman et al. (1997) and more recently by Versley (2006). However, as Passonneau (1997) points out, this is problematic for several reasons, the most important being that the measure by Vilain et al. (1995) disregards that a certain agreement between two annotators can also be given by chance. Thus, according to Passonneau, high mutual *recall* and *precision* of two annotations does not permit to conclude that the annotation is reliable.

To remedy the situation, Passonneau (2004) proposes Krippendorff's α (Krippendorff, 1980) as an alternative quantitative measure for agreement of annotations. α is similar to κ in that it factors out chance agreement. In the case of α, this is done by considering both the *observed* disagreement (D_o) and the *expected* disagreement (D_e). α is then calculated as follows:

$$\alpha = 1 - \frac{D_o}{D_e}$$

For α to be applicable, the annotation results have to be cast in the form of a reliability data matrix (Krippendorff, 1980), i.e. a table with one column for each item to be annotated (i.e. each markable) and one row for each annotator. Each table cell contains the annotation applied by the respective annotator for the respective markable. Passonneau (2004) suggests to identify the annotation for a given markable with the representation of the entire markable set that was formed by the markable being linked (transitively) to all other markables in the set. If a unique label is assigned to every markable partition observed in the entire annotation, this label can serve as the value in the reliability data matrix. In an annotation done by four annotators on six markables, e.g., the following different partitions might be observed: Partition 1 = {A, C}, Partition 2 = {A, C, D}, Partition 3 = {B, E}, Partition 4 = {D, F}. When distributed to the four annotators, the resulting reliability data matrix might look like Table 13.

In the above table, identical values within one column mean that the respective annotators agreed in their annotation. The first row e.g. states that annotators 1, 2, and 4

	A	B	C	D	E	F
1	Part 1	Part 3	Part 1	Part 4	Part 3	Part 4
2	Part 1	Part 3	Part 1	Part 4	Part 3	Part 4
3	Part 2	Part 3	Part 2	Part 2	Part 3	-
4	Part 1	Part 3	Part 1	Part 4	Part 3	Part 4

Table 13: Example reliability data matrix for six markables and four annotators.

agreed in assigning markables A and C to the same markable set (Partition 1), while annotator 2 created a markable set comprising markables A, C, and D (Partition 2).
On the basis of a reliability data matrix like that in Table 13, D_o is calculated by summing over the observed disagreements (cf. below). D_e, on the other hand, is calculated on the basis of the number of annotators and the number of markables alone, as the disagreement that would be expected if the annotation was attributable to mere chance.

An integral part of the definition of Krippendorff's α is the difference function δ. δ is a two-place function that returns for a given pair of observations (here: a pair of markable sets) a numerical value of the difference between both observations. The idea of Passonneau (2004) is to define this difference function in such a way that it does justice to the different degrees of dissimilarity that can hold between two sets A and B, i.e. identity, subsumption, overlap, and disjunction.[22] She proposes to use the following distances:

$$\delta_{AB} = \begin{cases} 0 & \text{if A = B, identity} \\ 1/3 & \text{if A} \subset \text{B or B} \subset \text{A, subsumption} \\ 2/3 & \text{if A} \cap \text{B} \neq \emptyset\text{, but A} \not\subseteq \text{B and B} \not\subseteq \text{A, overlap} \\ 1 & \text{if A} \cap \text{B} = \emptyset\text{, disjunction} \end{cases}$$

Passonneau (2004) states that when comparing two markable sets, care has to be taken to temporarily exclude from both sets the one markable whose classification is to be

[22] Poesio & Artstein (2005a) found that for their experiments, implementing δ on the basis of distance measures like Dice and Jaccard worked better. It has to be observed, however, that their annotation task is considerably different from the one employed in this thesis, as they allow annotators to mark arbitrary regions of text as antecedents. It is thus not clear whether the Dice- and Jaccard-based implementation of δ is superior here as well.

compared. For example, if annotator 1 created a markable set {A, B, C}, and annotator 2 created a markable set {A, X, Y}, according to the above definition, both sets overlap. However, when what is to be compared is the classification of markable A, this is wrong, since all markables except for A are different, and the intuitively correct result would be for both sets to be disjunct. Thus, Passonneau proposes to temporarily remove A from both sets prior to comparison, and to compare {B, C} and {X, Y}, which yields the correct result. Poesio & Artstein (2005b) point out that while this is correct for the calculation of the observed disagreement D_o, it is infeasible for the calculation of the expected disagreement D_e. This is because for the latter, no element can be removed since there is no element whose classification is currently compared. Poesio & Artstein (2005b) observe that this leads to inconsistencies because D_o and D_e are calculated on different bases.[23] Consequently, Poesio & Artstein (2005b) report two variants of α for their experiments. Under the *exclusive chain condition*, the current item is removed for the calculation of D_o as proposed by Passonneau, accepting the inconsistencies mentioned above. Under the *inclusive chain condition*, the current item is not removed, at the expense of a D_o that is unintuitively low, because cases of disjunction are counted as overlapping due to the common markable whose classification is to be compared.

According to Passonneau (2004), the application of α requires that all annotations contain the same set of markables. This condition is obviously not met in this thesis. As described above, all antecedents (except in cases where the antecedent was *it*, *this*, and *that*) were created by the annotators individually. Thus, we frequently encounter cases where one or more annotators created a markable that one or more other annotators did not create. According to Passonneau (2004), this causes the sets of markables to be incommensurate and α to be inapplicable. In the following, therefore, we will report α values computed on the *intersection* of the compared annotations, i.e. on the subset of those markables that can be found in the annotations of all four annotators. Obviously, this gives only a partial picture of actual disagreement because it ignores all cases where disagreement results from different or missing markables. In particular, this figure can only be interpreted in relation to how many of all distinct markables the annotators agreed upon at all. Therefore, we also provide the absolute size of the intersection as well as its relative size in proportion to all distinct markables created by all annotators.

[23] Artstein (p.c.) points out that Passonneau's method of calculating D_o even amounts to a modification of Krippendorff's α itself, which explicitly assumes D_o and D_e to be computed on the same basis.

Only a subset of the markables in each annotation is relevant for the determination of inter-annotator agreement on anaphora. This subset includes mainly two types of expressions: all non-pronominal markables, i.e. all markables manually created by the annotators as antecedents, and all instances of *it*, *this*, and *that* that have been assigned to a non-empty markable set. Table 14 contains figures for both of these subsets together (all), and for pronouns and non-pronouns individually. The second column in the table contains the cardinality of the union of all four annotators, i.e. the number of all distinct markables of the respective expression found in all four annotations. The third and fourth column contain the same figure for the intersection of these four data sets, both in absolute figures and in percentage of the previous figure. The fifth and sixth column contain the actual α values under the exclusive and the inclusive chain condition (cf. above) calculated on the markables in the intersection only.

	Expression	$\|1 \cup 2 \cup 3 \cup 4\|$	$\|1 \cap 2 \cap 3 \cap 4\|$		α (excl.)	α (incl.)
Bed017	all	397	109	27.46 %	.47	.57
	pronouns	173	94	54.34 %	.41	.52
	non-pronouns	224	15	6.70 %	.83	.84
Bmr001	all	619	195	31.50 %	.43	.51
	pronouns	312	179	57.37 %	.40	.48
	non-pronouns	307	16	5.21 %	.78	.78
Bns003	all	529	131	24.76 %	.45	.55
	pronouns	229	114	49.78 %	.41	.53
	non-pronouns	280	17	6.07 %	.70	.71
Bro004	all	703	142	20.20 %	.45	.55
	pronouns	317	126	39.75 %	.40	.51
	non-pronouns	386	16	4.15 %	.87	.87
Bro005	all	530	132	24.91 %	.52	.61
	pronouns	248	109	43.95 %	.44	.55
	non-pronouns	282	23	8.16 %	.87	.88

Table 14: Reliability (Krippendorff's α) for four annotators.

In Table 14, the figures in the first row (all) are the ones that are relevant for the *overall* agreement. In the five dialogs, the four annotators on average only agreed on the identification of markables in 27.77% of cases. Agreement (α) of the anaphora annotation in these five subsets ranges from $.43$ to $.52$ resp. $.51$ to $.61$. If only pronouns are considered (second row), the situation changes somewhat. Here, the annotators agreed on average in 49.04% of cases on a given pronoun to be referential.[24] Anaphora annotation agreement (α) among referential pronouns in the five subsets ranges from $.40$ to $.44$

[24] Pronoun markables were created automatically prior to the annotation, so the mere existence of a

resp. $.48$ to $.55$. Finally, if only non-pronominal antecedent expressions are considered (third row), the difficulty and high degree of subjectivity of the determination of (NP and VP) antecedents becomes apparent: On average, the annotators only agree on 6.06% of cases. In other words, more than 90% of all non-pronominal antecedents were identified by at most three out of four annotators. Among the antecedents that the annotators agree upon, however, the agreement of the anaphora annotation is considerable, ranging from $.70$ to $.87$ resp. $.71$ to $.88$. In general, the result of the analysis of the agreement supports the claim that the annotation of anaphoric relations in spoken dialog is very difficult and inherently ambiguous. In order for the data to be usable despite of this, data consolidation measures are required (cf. Chapter 4.2.4 below).

The reliability figures only give a partial, quantitative picture. In the following (Figure 9, page 84), we complement this with the qualitative analysis of a case of disagreement. We first show a fragment of a dialog, with the relevant expressions rendered in brackets and bold font for better visibility. The dialog fragment is shown in its original segmentation, which is the same format as that used in the display of the annotation tool. In the figure following the dialog fragment, the anaphoric relations are shown that were annotated by the four annotators. Expressions not participating in any anaphoric relations are left out of this figure. Each annotator is associated with a different line style.

Only two annotators (2 and 3) agreed in linking *that*$_1$ to the VP antecedent *to plan*, whereas the other two annotators selected *wanted* resp. *some Pareto optimal* as the antecedent. Note that annotator 4 identified only part of the NP antecedent, as the full NP would have been *some Pareto optimal thing*. For *that*$_2$, three annotators (2, 3 and 4) identified *that*$_1$ as the antecedent. The remaining annotator 1 linked *that*$_2$ to the VP antecedent *to plan*. Thus, three annotators (1, 2 and 3) identified an anaphoric relation between *that*$_2$ and *to plan*. One did so directly, the other two transitively via *that*$_1$. *that*$_3$ has only two different connections to other expressions, so only two of the four annotators identified it as anaphoric at all. Both annotators (1 and 4) link *that*$_3$ to *that*$_2$. Finally, the same two annotators linked *it* to *that*$_3$, while the other two selected *that*$_2$ as the antecedent for *it*. Thus, an anaphoric relation between *it* and *that*$_2$ was identified by all four annotators.

pronoun markable in all four annotations is not a sign of agreement. Instead, agreement is measured on the pronoun markables that the annotators added to a markable set.

ME010: Yeah. So, um,

ME010: you could,

ME010: from this, go on and say suppose there's a group of people traveling together

MN059: *outbreath*

ME010: and you **[wanted]** **[to plan]** something that somehow,

ME010: with **[some Pareto optimal]**

FE004: *laugh* **[That]**'s good. *laugh*

ME010: uh, *laugh* uh, thing for -

MN015: *laugh*

FN050: *laugh*

FE004: **[That]**'s definitely a job for artificial intelligence. *laugh*

MN015: *laugh*

ME010: uh, or -

MN015: Well **[that]**'s not - not even something humans - yeah. *laugh*

FE004: Except for humans can't really solve **[it]** either, so.

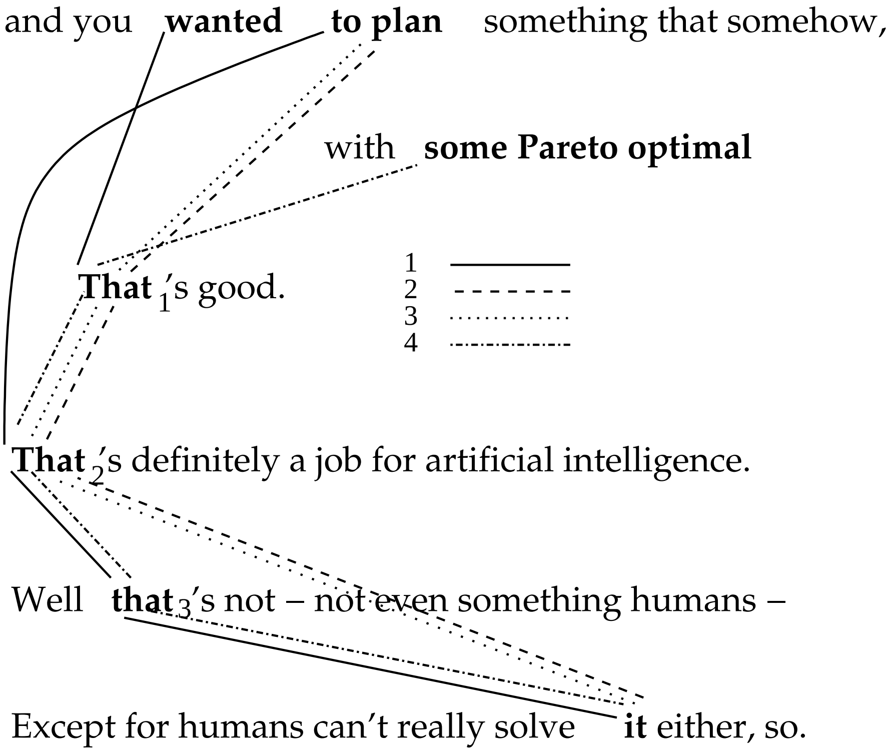

and you **wanted** **to plan** something that somehow,
with **some Pareto optimal**
That $_1$'s good.
1
2
3
4
That $_2$'s definitely a job for artificial intelligence.
Well **that** $_3$'s not – not even something humans –
Except for humans can't really solve **it** either, so.

Figure 9: A complex case of disagreement.

4.2.4 Automatic *Core* Data Set Generation

Just like for the classification task concerning *it*, cf. Chapter 4.1, we also needed a consolidated version of the anaphora data set. As mentioned above, the common method for obtaining this kind of data is to have the annotators manually create a gold standard version of their individual annotations. A tacit assumption that underlies the gold standard approach is that disagreements in annotations are mainly due to errors or misunderstandings on the part of one of the annotators, and that they simply have to be corrected. In view of the subjectivity of the annotation task, which is partly reflected in the low agreement even on markable *identification*, the manual creation of a consensus-based gold standard data set did not seem feasible. Note that this is in contrast to other anaphor annotation approaches in spoken dialog. Both Eckert & Strube (2000) and Byron (2003) create a consensus-based reconciled version of their annotations, and use this as the data basis for the application of their algorithms. However, neither Eckert & Strube (2000) nor Byron (2003) use really naive annotators. In addition, Eckert & Strube (2000) do the annotation incrementally, excluding contentious markables from subsequent annotation steps. Therefore, the data set that they eventually have to reconcile does no longer contain e.g. markables that are ambiguous between having an NP or VP referent.

As an alternative to a manually created gold standard data set based on only two annotators, we automatically created *core* data sets from the annotations of all four annotators by means of majority decisions. The rationale for this was that the more annotators arrived at the same interpretation independently of each other, the more plausible the respective interpretation would be.

The core data sets were generated by automatically collecting in each dialog those antecedent-anaphor pairs that at least n annotators identified independently of each other. The value for n ranged from two to four. The parameter n could be used to control which links should be allowed into the respective data set. With $n = 2$, the criterion was least restrictive, accepting any link that was identified by at least two annotators. With $n = 4$, on the other hand, the criterion was most restrictive, because a link was only accepted if it was identified by all four annotators. We assumed that in a difficult or ambiguous case, different annotators are less likely to agree independently of each other than in a more simple or clearcut case. Under this assumption, a high value for n could be used to exclude difficult and ambiguous cases from the data set.

The algorithm used for the generation of the core data sets (`Create_Core_Data`) is outlined in Algorithm 1. The algorithm is rather strict in that it selects a link only if all n annotators identified *exactly* the same pair of anaphor and immediate antecedent. In other words: Transitive links (like those specified by annotators 1 and 4 between *it* and *that*$_2$ in Figure 9 on page 84) are not considered. This restriction was necessary in order to prevent inconsistencies in the generated data.

The algorithm is greedy, i.e. it selects the first link that is found n times. This has a potential consequence if $n = 2$ and if an anaphor was linked to one antecedent by two annotators and to a different antecedent by the other two annotators. This situation holds for *it* in Figure 9 (page 84). Here, the decision on which of the two links is selected depends on the order in which the annotations by the four annotators are searched. In the example given, the link that will be selected is the one between *it* and *that*$_2$, because it was identified by annotators 2 and 3, and so this link is found twice before the one that was identified by annotators 1 and 4.

For each selected link, the algorithm stores the information that the respective antecedent has occurred as an antecedent and that the respective anaphor has occurred as an anaphor. This way, data inconsistencies in the core data sets are prevented, since for each value of n, each markable can appear at most once as an antecedent and at most once as an anaphor. This ensures that in cases where different annotators created conflicting links for the same markables (cf. above), only the first link is selected for the core data set.

Table 15 contains for each dialog the number of links in the respective core data set.

	$n = 2$	$n = 3$	$n = 4$
Bed017	116	62	28
Bmr001	229	132	69
Bns003	164	82	27
Bro004	212	111	32
Bro005	170	95	32
Σ	891	482	188

Table 15: No. of immediate links in three core data sets.

As was to be expected, the number of links in each data set decreases as the number of annotators that have to agree increases. Due to the way in which the data sets were generated, a set with a smaller n subsumes all other sets with a higher n for the same dialog.

Algorithm 1 Algorithm `Create_Core_Data`

```
for all Dialogs dialog in (Bed017, Bmr001, Bns003, Bro004, Bro005) do
  for $n = 2$ to $4$ do
    for all Annotators annotatorA in (1,2,3,4) do
      for all Coreference sets primarySet marked by annotatorA in dialog do
        for all Immediate ante-ana links primaryLink in primarySet do
          $found \leftarrow 1$
          for all Annotators annotatorB in (1,2,3,4) such that $annotatorB \neq annotatorA$ do
            for all Coreference sets secondarySet marked by annotatorB in dialog do
              if secondarySet exactly contains primaryLink then
                found++
              end if
              if $found = n$ then
                if primaryLink.ante has not occurred as antecedent then
                  if primaryLink.ana has not occurred as anaphor then
                    Add primaryLink to core data set $n$ for dialog
                    Store that primaryLink.ante has occurred as antecedent
                    Store that primaryLink.ana has occurred as anaphor
                  end if
                end if
              end if
            end for
          end for
        end for
      end for
    end for
  end for
end for
```

4.2.5 Annotated Corpus Analysis

In contrast to the data sets produced by e.g. Eckert & Strube (2000) or Byron (2003), our core data sets were not created manually, but automatically. The criterion for adding or not adding an anaphoric link to a set was not a qualitative one, e.g. plausibility based on human inspection, but a quantitative one, i.e. mere frequency of occurrence in a set of manual annotations. The underlying rationale, as already mentioned, is that a link is the more plausible the more often it was identified by individual annotators. This, however, can only serve as an approximation of plausibility. As a result, our core data sets are prone to two types of errors: First, they will contain some spurious or dubious links, which are caused by several annotators making the same mistake. Second, and more importantly, they will also lack some correct but more difficult links, on which the minimum number of annotators did not happen to agree. These facts have to be taken into account when interpreting frequency distributions or other descriptive measures. A high or low frequency of a phenomenon cannot simply be taken to reflect the 'real' distribution of this phenomenon. All that is known is the distribution in the artificially created, majority-based data subset. In particular, a low frequency of a phenomenon can have two reasons: Either the phenomenon is simply rare, or it is one of normal or even high frequency on which the agreement of the annotators just happens to be low. What can more safely be interpreted in the core data sets is the difference between the relative frequencies of a given phenomenon for different values of n. As mentioned above, links in a set for a bigger n are considered less difficult than those in a set for a smaller n. Under this assumption, the increase resp. decrease of a phenomenon with increasing n can be taken to indicate that the phenomenon is less resp. more difficult to identify.

In spite of their limitations, we argue that the core data sets described above represent a plausible approximation to the phenomenon of anaphoric reference in dialog, and that they contain a relevant subset of anaphoric links that are useful to resolve. In the following, we present an analysis of the core data sets based on simple frequency-based statistics.

4.2.5.1 Antecedent Type and Anaphor Frequencies We analyzed the distribution of antecedent types, i.e. whether the antecedent in an antecedent-anaphor pair is an NP, pronoun, VP, or other (adjective or nominalization). Table 16 details the distribution of antecedent types for every dialog and for all dialogs together in the respective core data set in absolute (top) and relative (bottom) figures.

	$n = 2$				$n = 3$				$n = 4$			
ante	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	43	46	22	5	23	32	7	0	11	16	1	0
	37.07	39.66	18.97	4.31	37.10	51.61	11.29	0.00	39.29	57.14	3.57	0.00
Bmr001	63	124	34	8	23	95	14	0	11	56	2	0
	27.51	54.15	14.85	3.49	17.42	71.97	10.61	0.00	15.94	81.16	2.90	0.00
Bns003	53	68	42	1	25	39	18	0	8	14	5	0
	32.32	41.46	25.61	0.61	30.49	47.56	21.95	0.00	29.63	51.85	18.52	0.00
Bro004	86	92	31	3	47	51	10	3	14	17	1	0
	40.57	43.40	14.62	1.42	42.34	45.95	9.01	2.70	43.75	53.13	3.13	0.00
Bro005	83	61	21	5	45	37	10	3	17	11	3	1
	48.82	35.88	12.35	2.94	47.37	38.95	10.53	3.16	53.13	34.38	9.38	3.13
Σ	328	391	150	22	163	254	59	6	61	114	12	1
	36.81	43.88	16.84	2.47	33.82	52.70	12.24	1.25	32.45	60.64	6.38	0.53

Table 16: Immediate antecedent types in three core data sets.

The figures in Table 16 show some clear trends: First, VP is the least frequent antecedent type in the core data sets (apart from OTHER, which is negligible). Even in the least restrictive core data set ($n = 2$), only 150 links (16.84%) are of type VP. The proportion gets smaller for the other two core data sets, dropping to 59 links (12.24%) for $n = 3$ and as little as twelve links (6.38%) for $n = 4$. NP antecedents are of medium frequency. Compared to VP antecedents, their rate is almost constant and drops only marginally with increasing n from 36.81% to 33.82% to 32.45%. Finally, the proportion of PRO antecedents increases with increasing n: For $n = 2$ it is already as high as 43.88%, and increases to 52.70% for $n = 3$ and to as much as 60.64% for $n = 4$. The drop in the proportion of VP antecedents with increasing n can be taken to indicate that anaphoric relations to VP antecedents are more difficult to identify than anaphoric relations to other types of antecedents. Similarly, the increase in the proportion of PRO antecedents shows that anaphoric relations to pronoun antecedents are more easy to identify.

We also analyzed the distribution of anaphors, i.e. whether the anaphor in an antecedent-anaphor pair is *it*, *this*, or *that*. Table 17 details the distribution of anaphors for every dialog and for all dialogs together in the respective core data set in absolute (top) and relative (bottom) figures. The figures for *its* are also included, but, as can be seen, they are negligible.

It can be seen that *this* is the least frequent anaphor in all three core data sets (apart from *its*). This finding is in line with Schiffman (1985) and Byron (2003).[25] The frequency of

[25] Schiffman (1985) reports that in her 31798 word corpus, there are only six tokens of anaphoric *this*.

	$n = 2$				$n = 3$				$n = 4$			
ana	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*
Bed017	48	4	22	42	32	3	9	18	18	0	4	6
	41.38	3.45	18.97	36.21	51.61	4.84	14.52	29.03	64.29	0.00	14.29	21.43
Bmr001	156	0	11	62	99	0	4	29	59	0	1	9
	68.12	0.00	4.80	27.07	75.00	0.00	3.03	21.97	85.51	0.00	1.45	13.04
Bns003	96	1	20	47	48	1	9	24	19	0	0	8
	58.54	0.61	12.20	28.66	58.54	1.22	10.98	29.27	70.37	0.00	0.00	29.63
Bro004	142	0	20	50	84	0	6	21	25	0	1	6
	66.98	0.00	9.43	23.59	75.68	0.00	5.41	18.92	78.13	0.00	3.13	18.75
Bro005	95	1	26	48	53	1	11	30	17	0	0	15
	55.88	0.59	15.29	28.24	55.79	1.05	11.58	31.58	53.13	0.00	0.00	46.86
Σ	537	6	99	249	316	5	39	122	138	0	6	44
	60.27	0.67	11.11	27.95	65.56	1.04	8.10	25.31	73.40	0.00	3.19	23.40

Table 17: Anaphor types in three core data sets.

this drops from 99 (11.11%) for $n = 2$ to 39 (8.1%) for $n = 3$ to six (3.19%) for $n = 4$. Like for the figures in Table 16, the decrease of relative frequency with increasing n can be interpreted as a tendency of *this* to be more difficult to interpret by the annotators than other pronouns like e.g. *it*. The pronoun *that* is of medium frequency in the core data sets. Its relative frequency is 27.95% for $n = 2$ and drops slightly to 25.31% for $n = 3$ and to 23.4% for $n = 4$. The pronoun *it* is the most frequent anaphor in the core data sets. Also, its rate increases with increasing n from 60.27% for $n = 2$ to 65.56% for $n = 3$ to 73.4% for $n = 4$, indicating a tendency to be more easily interpreted by the annotators. Schiffman (1985) also finds *it* to be more frequent than *that* (838 vs. 582 instances), while in Byron (2003), *that* is more frequent than *it* (162 vs. 122 instances).

4.2.5.2 Anaphor-Antecedent Pair Distribution Having considered the distributions of different antecedent types and anaphors individually, we were also interested in correlations between both. In the following, we present three pairs of tables (Tables 18 to 23), one pair for each n. The first table in each pair contains the absolute and relative frequencies for the anaphor type, broken down according to the type of immediate antecedent. The other table contains the same number for the immediate antecedent type, broken down according to the type of anaphor. The values in the highlighted table cells are the highest values for the resp. dialog and for the resp. anaphor-antecedent pair. E.g., the top leftmost cell in Table 19 is highlighted because in dialog Bed017 most antecedents of the anaphor *it* (47.92%) are of type NP.

The 10420 word corpus of Byron (2003) contains only five instances of anaphoric *this*.

It was mentioned in Chapter 2.3 that previous research in the functions of *it*, *this*, and *that* showed a couple of rather stable principles. One of these was the preference of *that* to be discourse-deictic, resp. of discourse-deictic reference to be realized preferrably by means of *that*. In our annotation, discourse deixis is defined as anaphoric reference to VP antecedents, so if the principle holds for our data as well, it should be notable in the correlation between *that* and VP antecedents.

In the core data set for $n = 2$ (Table 19, page 92, bottom) it can be seen that in total the anaphor *that* is more often associated with a VP antecedent (36.95%) than with any other single antecedent type. On the level of the individual dialogs, this is true for only three of five dialogs. In dialog Bmr001, the number of PRO antecedents for *that* is minimally higher than the number of VP antecedents. In dialog Bro005, as many as 50% of antecedents for *that* are of type NP. The inverse correlation, i.e. the preference of a VP antecedent to be referred to anaphorically by *that*, can be seen in Table 18, page 92, top. Here, the situation is much clearer, as in total more than 60% of all VP antecedents in the core level data set for $n = 2$ are referred to anaphorically by *that*. Also, this trend is valid without exception in all five individual dialogs. VP antecedents are more often associated with *that* than with all other anaphor types taken together.

For $n = 3$ (Table 21, page 93, bottom), there is no preference for *that* to refer to any particular type of antecedent. Both PRO and VP antecedents occur in 33.61% of all anaphor-antecedent pairs in which *that* is the anaphor, and with 31.15%, NP antecedents are only slightly less frequent. The inverse relation, on the other hand, is even more pronounced than for $n = 2$, as can be seen in Table 20, page 93, top. Almost 70% of all VP antecedents in the core data set for $n = 3$ are anaphorically referred to by *that*.

Finally, for $n = 4$, the antecedent preference for *that* (Table 23, page 94, bottom) is shifted to NP, which accounts for 45.46% of all antecedents for *that*. The inverse relation, on the other hand, remains stable, as can be seen in Table 22, page 94, top. Nine out of twelve VP antecedents in the core data set for $n = 4$ (75%) are anaphorically referred to by *that*. Our core data sets thus corroborate the finding of other researchers that for discourse-deictic reference, *that* is preferred over *it*.

4.2.5.3 Distances Another descriptive measure for anaphoric links is the average distance between anaphors and their antecedents. Table 24 (page 96) contains for each core data set the average distances (and standard deviations) in words and in seconds, broken down according to the type of antecedent, i.e. NP, PRO, nominal (NP + PRO), and

ante	NP				PRO				VP				OTHER			
ana	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*
Bed017	**23** **53.49**	3 6.98	6 13.95	11 25.58	**22** **47.83**	1 2.17	12 26.09	11 23.91	3 13.64	0 0.00	4 18.18	**15** **68.18**	0 0.00	0 0.00	0 0.00	**5** **100.00**
Bmr001	**44** **69.84**	0 0.00	5 7.94	14 22.22	**98** **79.03**	0 0.00	4 3.23	22 17.74	12 35.29	0 0.00	1 2.94	**21** **61.77**	2 25.00	0 0.00	1 12.50	**5** **62.50**
Bns003	**35** **66.04**	0 0.00	7 13.21	11 20.76	**52** **76.47**	1 1.47	6 8.82	9 13.24	9 21.43	0 0.00	7 16.67	**26** **61.91**	0 0.00	0 0.00	0 0.00	**1** **100.00**
Bro004	**62** **72.09**	0 0.00	8 9.30	16 18.61	**72** **78.26**	0 0.00	6 6.52	14 15.22	7 22.58	0 0.00	5 16.13	**19** **61.29**	**1** **33.33**	0 0.00	**1** **33.33**	**1** **33.33**
Bro005	**51** **61.45**	1 1.21	7 8.43	24 28.92	**37** **60.66**	0 0.00	13 21.31	11 18.03	5 23.81	0 0.00	5 23.81	**11** **52.38**	**2** **40.00**	0 0.00	1 20.00	**2** **40.00**
Σ	**215** **65.55**	4 1.22	33 10.06	76 23.17	**281** **71.87**	2 0.51	41 10.49	67 17.14	36 24.00	0 0.00	22 14.67	**92** **61.33**	5 22.73	0 0.00	3 13.64	**14** **63.64**

Table 18: Anaphor type distribution in core data set for $n = 2$.

ana	*it*				*its*				*this*				*that*			
ante	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	**23** **47.92**	22 45.83	3 6.25	0 0.00	**3** **75.00**	1 25.00	0 0.00	0 0.00	6 27.27	**12** **54.55**	4 18.18	0 0.00	11 26.19	11 26.19	**15** **35.71**	5 11.91
Bmr001	44 28.21	**98** **62.82**	12 7.69	2 1.28	0 0.00	0 0.00	0 0.00	0 0.00	**5** **45.46**	4 36.36	1 9.09	1 9.09	14 22.58	**22** **35.48**	21 33.87	5 8.07
Bns003	35 36.46	**52** **54.17**	9 9.38	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	**7** **35.00**	6 30.00	**7** **35.00**	0 0.00	11 23.40	9 19.15	**26** **55.32**	1 2.13
Bro004	62 43.66	**72** **50.70**	7 4.93	1 0.70	0 0.00	0 0.00	0 0.00	0 0.00	**8** **40.00**	6 30.00	5 25.00	1 5.00	16 32.00	14 28.00	**19** **38.00**	1 2.00
Bro005	**51** **53.68**	37 38.95	5 5.26	2 2.11	**1** **100.00**	0 0.00	0 0.00	0 0.00	7 26.92	**13** **50.00**	5 19.23	1 3.85	**24** **50.00**	11 22.92	11 22.92	2 4.17
Σ	215 40.04	**281** **52.33**	36 6.70	5 0.93	**4** **66.67**	2 33.33	0 0.00	0 0.00	33 33.33	**41** **41.41**	22 22.22	3 3.03	76 30.52	67 26.91	**92** **36.95**	14 5.62

Table 19: Immediate antecedent type distribution in core data set for $n = 2$.

ante	NP				PRO				VP				OTHER			
ana	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*
Bed017	**14** **60.87**	2 8.70	0 0.00	7 30.44	**18** **56.25**	1 3.13	7 21.88	6 18.75	0 0.00	0 0.00	2 28.57	**5** **71.43**	0 0.00	0 0.00	0 0.00	0 0.00
Bmr001	**18** **78.26**	0 0.00	0 0.00	5 21.74	**76** **80.00**	0 0.00	4 4.21	15 15.79	5 35.71	0 0.00	0 0.00	**9** **64.29**	0 0.00	0 0.00	0 0.00	0 0.00
Bns003	**18** **72.00**	0 0.00	3 12.00	4 16.00	**28** **71.80**	1 2.56	4 10.26	6 15.39	2 11.11	0 0.00	2 11.11	**14** **77.78**	0 0.00	0 0.00	0 0.00	0 0.00
Bro004	**38** **80.85**	0 0.00	1 2.13	8 17.02	**43** **84.31**	0 0.00	2 3.92	6 11.77	2 20.00	0 0.00	2 20.00	**6** **60.00**	**1** **33.33**	0 0.00	**1** **33.33**	**1** **33.33**
Bro005	**28** **62.22**	1 2.22	2 4.44	14 31.11	**22** **59.46**	0 0.00	7 18.92	8 21.62	2 20.00	0 0.00	1 10.00	**7** **70.00**	**1** **33.33**	0 0.00	**1** **33.33**	**1** **33.33**
Σ	**116** **71.17**	3 1.84	6 3.68	38 23.31	**187** **73.62**	2 0.79	24 9.45	41 16.14	11 18.64	0 0.00	7 11.86	**41** **69.49**	**2** **33.33**	0 0.00	**2** **33.33**	**2** **33.33**

Table 20: Anaphor type distribution in core data set for $n = 3$.

ana	*it*				*its*				*this*				*that*			
ante	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	14 43.75	**18** **56.25**	0 0.00	0 0.00	**2** **66.67**	1 33.33	0 0.00	0 0.00	0 0.00	**7** **77.78**	2 22.22	0 0.00	**7** **38.89**	6 33.33	5 27.78	0 0.00
Bmr001	18 18.18	**76** **76.77**	5 5.05	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**4** **100.00**	0 0.00	0 0.00	5 17.24	**15** **51.72**	9 31.03	0 0.00
Bns003	18 37.50	**28** **58.33**	2 4.17	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	3 33.33	**4** **44.44**	2 22.22	0 0.00	4 16.67	6 25.00	**14** **58.33**	0 0.00
Bro004	38 45.24	**43** **51.19**	2 2.28	1 1.19	0 0.00	0 0.00	0 0.00	0 0.00	1 16.67	**2** **33.33**	**2** **33.33**	1 16.67	**8** **38.10**	6 28.57	6 28.57	1 4.76
Bro005	**28** **52.83**	22 41.51	2 3.77	1 1.88	**1** **100.00**	0 0.00	0 0.00	0 0.00	2 18.18	**7** **63.64**	1 9.09	1 9.09	**14** **46.67**	8 26.67	7 23.33	1 3.33
Σ	116 36.71	**187** **59.18**	11 3.48	2 0.63	**3** **60.00**	2 40.00	0 0.00	0 0.00	6 15.39	**24** **61.54**	7 17.95	2 5.13	38 31.15	**41** **33.61**	**41** **33.61**	2 1.64

Table 21: Immediate antecedent type distribution in core data set for $n = 3$.

ante	NP				PRO				VP				OTHER			
ana	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*	*it*	*its*	*this*	*that*
Bed017	**7** **63.64**	0 0.00	0 0.00	4 36.36	**11** **68.75**	0 0.00	3 18.75	2 12.50	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
Bmr001	**10** **90.91**	0 0.00	0 0.00	1 9.09	**47** **83.93**	0 0.00	1 1.79	8 14.29	**2** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00
Bns003	**6** **75.00**	0 0.00	0 0.00	2 25.00	**13** **92.86**	0 0.00	0 0.00	1 7.14	0 0.00	0 0.00	0 0.00	**5** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00
Bro004	**8** **57.14**	0 0.00	1 7.14	5 35.71	**17** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	0 0.00	0 0.00
Bro005	**9** **52.94**	0 0.00	0 0.00	8 47.06	**8** **72.73**	0 0.00	0 0.00	3 27.27	0 0.00	0 0.00	0 0.00	**3** **100.00**	0 0.00	0 0.00	0 0.00	**1** **100.00**
Σ	**40** **65.57**	0 0.00	1 1.64	20 32.79	**96** **84.21**	0 0.00	4 3.51	14 12.28	2 16.67	0 0.00	1 8.33	**9** **75.00**	0 0.00	0 0.00	0 0.00	**1** **100.00**

Table 22: Anaphor type distribution in core data set for $n = 4$.

ana	*it*				*its*				*this*				*that*			
ante	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	7 38.89	**11** **61.11**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**3** **75.00**	1 25.00	0 0.00	**4** **66.67**	2 33.33	0 0.00	0 0.00
Bmr001	10 16.95	**47** **79.66**	2 3.39	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	1 11.11	**8** **88.89**	0 0.00	0 0.00
Bns003	6 31.58	**13** **68.42**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	2 25.00	1 12.50	**5** **62.50**	0 0.00
Bro004	8 32.00	**17** **68.00**	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**1** **100.00**	0 0.00	0 0.00	0 0.00	**5** **83.33**	0 0.00	1 16.67	0 0.00
Bro005	**9** **52.94**	8 47.06	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	**8** **53.33**	3 20.00	3 20.00	1 6.67
Σ	40 28.99	**96** **69.57**	2 1.45	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	1 16.67	**4** **66.67**	1 16.67	0 0.00	**20** **45.46**	14 31.82	9 20.46	1 2.27

Table 23: Immediate antecedent type distribution in core data set for $n = 4$.

VP. For NP and PRO antecedents, distances are calculated from the rightmost word of the antecedent. For VP antecedents, two distinct ways of distance calculation were used. The first (VP) uses the rightmost word of the VP antecedent, i.e. of the verb, while the second (VP$_{phrase}$) uses the rightmost word of the entire phrase of which the antecedent is the head. As a result, the values in the column VP$_{phrase}$ are consistently smaller than those in the column VP, because the end of the phrase is closer to the anaphor than the head. The values in the column VP$_{phrase}$ are intended to capture the temporal distance between a discourse-deictic anaphor and the mention of its VP referent. We assume that this mention is not complete until the corresponding phrase (e.g. the proposition) has been completed. Since the ICSI Meeting Corpus does not contain word-level timing information, all temporal distances were calculated on the basis of a simple forced alignment (cf. Chapter 6.1.3). The main finding of the analysis of average distances is that both the average word and the average temporal distance decrease with increasing n. Apart from this, the average distances and standard deviations provide a useful empirical basis for estimating the antecedent search depth parameters for automatic resolution (cf. Chapter 7.2.1).

4.2.5.4 Anaphoric Chains While the link-based format is convenient for the calculation of local, pair-related phenomena like distance, we are mainly interested in the anaphoric chains that sequences of these links give rise to. Each chain represents an instance of a chain-initial mention and one or more pronominal re-mentions. It has to be noted, however, that there is no one-to-one relation between these chains and full-blown coreference chains (apart from the fact that, as described above, not all annotated relations qualify as coreference). One important difference is that, due to our method of annotation (Chapter 4.2.2), non-pronominal anaphoric expressions (in particular definite NPs) are not linked to their antecedents. Only anaphoric instances of *it*, *this*, and *that* are linked to their antecedents. As a result, longer chains including non-pronominal anaphors are split into smaller chains, each beginning with a non-pronominal expression.

The link-based core data sets were converted into anaphoric chains by transitively following all immediate links and putting the respective expressions into the same markable set. Using the discourse ordering of markables, each set could then be represented as an anaphoric chain. The number and lengths of the resulting anaphoric chains for each core data set can be found in Tables 25 to 27, broken down according to the type

		NP				PRO				Nominal (NP + PRO)				VP				VP_{phrase}			
		words		secs.		words		secs.		words		secs.		words		secs.		words		secs.	
		∅	σ	∅	σ	∅	σ	∅	σ	∅	σ	∅	σ	∅	σ	∅	σ	∅	σ	∅	σ
All	2	13.19	26.00	5.37	8.32	16.38	23.73	5.35	8.09	15.17	23.37	5.36	7.62	14.81	17.66	5.17	6.24	6.68	13.03	2.62	4.30
	3	10.14	13.30	4.31	4.79	14.62	20.14	4.57	6.51	12.87	16.65	4.47	5.39	14.32	15.76	5.15	5.85	6.47	10.70	2.69	3.74
	4	9.62	8.68	4.08	4.01	10.64	9.16	3.08	2.69	10.29	7.45	3.43	2.64	10.83	12.75	3.98	5.06	3.33	5.80	1.45	2.36

Table 24: Avg. distances (∅) and standard deviations (σ) of antecedent and anaphor in three core data sets.

of chain-initial antecedent. The lower part of each table contains the summed values for the five individual dialogs, including (in the rightmost and leftmost columns) the relative frequency for the respective type. The rightmost column in each table contains the percentage of chains with a particular antecedent type, relative to all chains in the core data set. For example, for $n = 2$ (Table 25, page 98), 26.64% of all chains have a VP antecedent. For $n = 3$ and $n = 4$, the respective percentages are 17.20% and 8.28%. There is also a considerable number of chains in which the chain-initial antecedent is itself a pronoun. These chains constitute 11.90 % (for $n = 2$) resp. 33.53% (for $n = 3$) resp. 48.97% (for $n = 4$) of all anaphoric chains in the core data sets. Note that there are two reasons why an anaphoric chain with a pronominal chain-initial antecedent can exist. Either the chain-initial pronoun is *vague*, i.e. it does not have any identifiable antecedent at all, or it is anaphoric, but the antecedent failed to be identified by the required number of annotators. Both types of chains have in common that – due to the lack of a non-pronominal antecedent – their automatic resolution does not contribute any information in a setting in which resolved anaphors are to be substituted with their antecedents (cf. Chapter 7.1).

The leftmost column in each table's lower part contains for each type of chain-initial antecedent and for all chains together the percentage of anaphoric chains of a particular length. For example: For $n = 2$ (Table 25, page 98), 68.21% of all anaphoric chains consist of two elements, i.e. an antecedent and one anaphor, only. For $n = 3$ and $n = 4$, the respective percentages are 76.09% and 78.62%. The percentage of two-element chains among the chains with an initial NP antecedent is 67.38 % (for $n = 2$) and 76.07% (for $n = 3$) and 81.97% (for $n = 4$), respectively. For comparison: Schiffman (1985) reports that of 101 chains with an initial NP antecedent, 59 (=58%) consist of an antecedent and one anaphor only. The frequency of two-element chains vs. chains with more than two elements provides important information for choosing an appropriate automatic anaphor resolution strategy (cf. Chapter 7.2.9).

4.3 Chapter Summary

This chapter described in detail the two annotation experiments that were performed in order to collect the data that serve as the empirical basis for this thesis. For both experiments, care was taken to avoid any form of bias to interfere with the annotation process. Therefore, all experiments were performed by naive annotators who were not in any other way involved in the project. Another methodological requirement was that

		2	3	4	5	6	7	8	9	10	11	12	total
Bed017	NP	30	7	3	2	1	-	-	-	-	-	-	43
	PRO	6	2	-	-	1	-	-	-	-	-	-	9
	VP	19	1	-	1	-	1	-	-	-	-	-	22
	OTHER	2	1	-	-	-	-	-	-	-	-	-	3
	all	57	11	3	3	2	1	-	-	-	-	-	77
Bmr001	NP	37	13	5	2	2	2	1	-	-	-	1	63
	PRO	7	3	1	-	2	1	1	-	-	-	-	15
	VP	22	8	2	-	-	1	-	1	-	-	-	34
	OTHER	6	-	-	-	-	-	-	-	-	-	-	6
	all	72	24	8	2	4	4	2	1	-	-	1	118
Bns003	NP	33	8	10	2	-	-	-	-	-	-	-	53
	PRO	5	2	1	1	-	-	-	-	-	-	-	9
	VP	31	7	2	2	-	-	-	-	-	-	-	42
	OTHER	-	1	-	-	-	-	-	-	-	-	-	1
	all	69	18	13	5	-	-	-	-	-	-	-	105
Bro004	NP	57	19	6	3	-	1	-	-	-	-	-	86
	PRO	13	7	2	-	-	-	-	-	-	-	-	22
	VP	23	5	1	2	-	-	-	-	-	-	-	31
	OTHER	2	1	-	-	-	-	-	-	-	-	-	3
	all	95	32	9	5	-	1	-	-	-	-	-	142
Bro005	NP	64	15	3	-	1	-	-	-	-	-	-	83
	PRO	9	2	-	1	-	-	-	-	-	-	-	12
	VP	13	5	1	1	-	-	-	-	-	1	-	21
	OTHER	5	-	-	-	-	-	-	-	-	-	-	5
	all	91	22	4	2	1	-	-	-	-	1	-	121
Σ	NP	221 67.38	62	27	9	4	3	1	-	-	-	1	328 58.26
	PRO	40 59.70	16	4	2	3	1	1	-	-	-	-	67 11.90
	VP	108 72.00	26	6	6	-	2	-	1	-	1	-	150 26.64
	OTHER	15 83.33	3	-	-	-	-	-	-	-	-	-	18 3.20
	all	384 68.21	107	37	17	7	6	2	1	-	1	1	563 100.00

Table 25: Anaphoric chain statistics in core data set for $n = 2$.

		2	3	4	5	6	7	8	9	10	11	12	total
Bed017	NP	17	3	2	-	1	-	-	-	-	-	-	23
	PRO	14	-	2	-	-	-	-	-	-	-	-	16
	VP	6	1	-	-	-	-	-	-	-	-	-	7
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	37	4	4	-	1	-	-	-	-	-	-	46
Bmr001	NP	14	4	1	1	1	1	-	-	-	1	-	23
	PRO	19	9	2	2	1	-	1	-	-	-	-	34
	VP	9	5	-	-	-	-	-	-	-	-	-	14
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	42	18	3	3	2	1	1	-	-	1	-	71
Bns003	NP	18	3	3	1	-	-	-	-	-	-	-	25
	PRO	18	1	1	-	-	-	-	-	-	-	-	20
	VP	14	4	-	-	-	-	-	-	-	-	-	18
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	50	8	4	1	-	-	-	-	-	-	-	63
Bro004	NP	38	5	3	1	-	-	-	-	-	-	-	47
	PRO	21	4	-	1	-	-	-	-	-	-	-	26
	VP	8	1	1	-	-	-	-	-	-	-	-	10
	OTHER	2	1	-	-	-	-	-	-	-	-	-	3
	all	69	11	4	2	-	-	-	-	-	-	-	86
Bro005	NP	37	7	1	-	-	-	-	-	-	-	-	45
	PRO	15	3	1	-	-	-	-	-	-	-	-	19
	VP	8	1	-	1	-	-	-	-	-	-	-	10
	OTHER	3	-	-	-	-	-	-	-	-	-	-	3
	all	63	11	2	1	-	-	-	-	-	-	-	77
Σ	NP	124 76.07	22	10	3	2	1	-	-	-	1	-	163 47.52
	PRO	87 75.65	17	6	3	1	-	1	-	-	-	-	115 33.53
	VP	45 76.27	12	1	1	-	-	-	-	-	-	-	59 17.20
	OTHER	5 83.33	1	-	-	-	-	-	-	-	-	-	6 1.75
	all	261 76.09	52	17	7	3	1	1	-	-	1	-	343 100.00

Table 26: Anaphoric chain statistics in core data set for $n = 3$.

		2	3	4	5	6	7	8	9	10	11	12	total
	NP	10	1	-	-	-	-	-	-	-	-	-	11
	PRO	9	3	-	-	-	-	-	-	-	-	-	12
Bed017	VP	1	-	-	-	-	-	-	-	-	-	-	1
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	20	4	-	-	-	-	-	-	-	-	-	24
	NP	7	2	1	1	-	-	-	-	-	-	-	11
	PRO	23	4	3	2	-	-	-	-	-	-	-	32
Bmr001	VP	1	1	-	-	-	-	-	-	-	-	-	2
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	31	7	4	3	-	-	-	-	-	-	-	45
	NP	4	2	2	-	-	-	-	-	-	-	-	8
	PRO	6	1	-	-	-	-	-	-	-	-	-	7
Bns003	VP	5	-	-	-	-	-	-	-	-	-	-	5
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	15	3	2	-	-	-	-	-	-	-	-	20
	NP	13	1	-	-	-	-	-	-	-	-	-	14
	PRO	8	4	-	-	-	-	-	-	-	-	-	12
Bro004	VP	1	-	-	-	-	-	-	-	-	-	-	1
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	22	5	-	-	-	-	-	-	-	-	-	27
	NP	16	1	-	-	-	-	-	-	-	-	-	17
	PRO	6	2	-	-	-	-	-	-	-	-	-	8
Bro005	VP	3	-	-	-	-	-	-	-	-	-	-	3
	OTHER	1	-	-	-	-	-	-	-	-	-	-	1
	all	26	3	-	-	-	-	-	-	-	-	-	29
	NP	50 81.97	7	3	1	-	-	-	-	-	-	-	61 42.07
	PRO	52 73.24	14	3	2	-	-	-	-	-	-	-	71 48.97
Σ	VP	11 91.67	1	-	-	-	-	-	-	-	-	-	12 8.28
	OTHER	1 100.00	-	-	-	-	-	-	-	-	-	-	1 0.69
	all	114 78.62	22	6	3	-	-	-	-	-	-	-	145 100.00

Table 27: Anaphoric chain statistics in core data set for $n = 4$.

we wanted to apply appropriate measures to control the inter-annotator agreement and thus the reliability of each annotation task.

The first experiment (performed by two annotators) dealt with the classification of *it*, *this*, and *that* as belonging to one of five categories. The major aim was the collection of data for building a component for the automatic detection of non-referential *it*. The reliability of this annotation, measured in κ, turned out to be below a minimum threshold for at least some categories. By conflating those categories that are equivalent with respect to the distinction *referential* vs. *non-referential*, the κ values could be considerably improved. In a final step, the two annotators created a consensus-based gold standard variant.

The second annotation experiment (performed by four annotators) dealt with the identification of antecedents for anaphoric instances of *it*, *this*, and *that*. Like for the first experiment, the major aim was the collection of training and test data, this time for the development of a system for automatic anaphora resolution. This practical application defined a couple of requirements on the annotated data. The most important of these requirements were that

1. for each anaphoric (incl. discourse-deictic) pronoun, there had to be exactly one antecedent in the preceeding dialog, and
2. this antecedent must be identifiable fully automatically.

A review of some existing anaphora and coreference annotation schemes and manuals showed that most of them were not applicable for the present task. Some were too abstract resp. too demanding, which made them inappropriate for naive annotators. Others were simpler and more usable, but failed to meet the requirements, especially that of automatic antecedent identifiability. As a result, a new and very simple scheme was employed. One of its main characteristics was that it used a minimalist definition of VP antecedents (antecedents for discourse-deictic pronouns), viz. the finite or infinite verb, which in the context of this thesis is assumed to provide sufficient information for the identification of the larger unit (sentence or clause). The reliability of this second annotation, measured in a variant of Krippendorff's α, turned out to be very low. In fact, major disagreements could already be found for antecedent *identification*. In view of these results, the creation of a consensus-based gold standard annotation (like for the first annotation) did not seem feasible. Instead, consolidated versions of the annotations (so-called *core* data sets) were created automatically by means of majority decisions. The assumption underlying this approach was that an anaphoric link is the more plausible the more annotators identify it independently of each other. By using this quantitative criterion, it was possible to create meaningful data sets for our anaphora resolution experiments.

5 State of the Art in Spoken Dialog Pronoun Resolution

In this chapter we outline the current state of the art in spoken dialog pronoun resolution. So far, the amount of work that has been done on pronoun resolution in written text clearly outweighs that which has been done for spoken dialog. Therefore, in Chapter 5.1 we begin by rewieving some of these approaches. It is not the aim of Chapter 5.1 to give a complete overview of existing attempts to pronoun resolution in written text (for this, see e.g. Mitkov (2002)). Rather, we want to illustrate

- the 'historical' development of computational approaches to the task,
- the basic methodological paradigms, and
- the high degree of optimization and specialization that has already been reached.

For this reason, the order of Chapter 5.1 is roughly chronological. That chapter will then serve as a background for Chapters 5.2 and 5.3, which deal with unimplemented and implemented approaches to spoken dialog pronoun resolution, respectively.

In Chapter 5.2, it will become clear that the theoretical machinery that is required for processing spoken dialog is considerably different and more sophisticated than what is required for written text. One main point is the distinction between monologic text vs. multi-speaker dialog and the resulting surface-structural differences (linear sequence of sentences vs. potentially overlapping turns). This is of particular importance for Byron & Stent (1998) (Chapter 5.2.1). Another point has to do with dialog-specific referential categories like discourse deixis and vague reference (see Chapter 2.2.3 and 2.2.4, respectively). As stressed by Eckert & Strube (2000) (Chapter 5.2.3), discourse deixis in particular requires a discourse-structural analysis in terms of (in their case) dialog acts in order to be able to provide non-NP antecedent candidates.

In Chapter 5.3, then, two systems will be described which represent the current state of the art of implemented spoken dialog pronoun resolution. Different as the systems are, they are similar in that they make a couple of assumptions which render them both inapplicable in a setting as realistic as the one of the present thesis.

5.1 Pronoun Resolution in Written Text vs. Spoken Dialog

Pronoun (or general coreference) resolution in written (mainly newspaper) text is by now a well-established discipline in computational linguistics. Most existing approaches can be characterized according to which methodological paradigm they use and how they model the resolution process. One major distinction can be drawn between **symbolic** or **rule-based** approaches on the one and probabilistic **classification-based** approaches on the other hand. Approaches of the former type rely on hand-crafted rules or heuristics, while those of the second type are based on statistical models acquired automatically from annotated corpora by means of machine learning. A second major distinction can be drawn between approaches that model pronoun resolution as the mapping of **anaphors to antecedents** and those that model the process as the mapping of **anaphors to referents**. Both approaches have come to be known as mention-pair approach vs. entity-mention approach (Luo et al., 2004). In the mention-pair approach, what is being searched is the whole set of previous mentions. These are treated as if they were independent, i.e. regardless of the possibility that some of them are themselves coreferent. As a result, the mention-pair approach does not make full use of all available information, a point that has often been criticized by the advocates of the mention-entity approach. This approach, in contrast, tries to find an anaphor's referent (i.e. a discourse entity). For practical reasons, these are commonly represented as the cumulation of all previous mentions of the same referent. This has the advantage that the choice of a referent can be made on the basis of *all* of its mentions simultaneously, thus making better use of the available information. On the downside, however, the incremental nature of the entity-mention approach makes it more prone to error propagation, which can quickly lead to classifier deterioration.

Hobbs (1978)'s 'Naive Algorithm' is one of the earliest attempts towards algorithmic resolution of pronouns. It is also an example of a rule-based mention-pair approach. The algorithm is naive in that it mainly consists of a search through a syntactic parse tree. More specifically, when a pronoun is encountered in a sentence, the parse tree is traversed in a particular order, and every NP that matches the number and gender features of the pronoun is proposed as an antecedent candidate. If no candidate is found in the current sentence, the parse trees of previous sentences are traversed in a similar fashion. Hobbs (1978) reports that the manual application of the algorithm correctly resolves 88.3% of 300 pronouns from three different genres. This is impressive given the fact that the algorithm relies on number and gender agreement and the correct search order alone. On the other hand, however, this search order can only be employed if a complete and correct syntactic analysis is available, which severely limits the practical applicability of the algorithm.

Brennan et al. (1987) is one of the first implemented approaches to employ Centering for the resolution of pronouns. Centering (Grosz et al., 1995) is a formal framework which relates the choice of a particular referential form to the local coherence of the discourse. When viewed as a production model, it predicts, given a particular referent, which form of referring expression should be used in order to create or maintain local coherence. When viewed as a resolution model, it predicts, given a particular referring expression, which of a list of potential referents is the most probable one. This prediction is based on the assumption that the producer of a referring expression (i.e. the speaker/writer) strives for coherence when choosing the form of the expression, and that the receiver (i.e. the listener/reader) utilizes this knowledge when interpreting the utterance.

An important data structure in Centering is *Cf*, the list of so-called *forward-looking centers*. There is one such list for each utterance, and it contains all discourse entities (i.e. referents) that appear in the utterance. The *Cf* is ranked, and the ranking order most commonly used (Kehler, 1997) is Subject > Direct Object > Indirect Object > Other. The highest-ranked discourse entity in a particular *Cf* is referred to as that *Cf*'s *preferred center*, or *Cp*. Equally important is the *Cb*, the so-called *backward-looking center*. In the intersection of the *Cf*s of the current and the previous utterance, the *Cb* is the single discourse entity that is ranked highest according to its grammatical function.

The degree of local coherence is expressed in terms of how disruptive the transition between two adjacent utterances U_{n-1} and U_n is with respect to the previous and the current preferred center ($Cp(U_{n-1})$ vs. $Cp(U_n)$) resp. the previous and the current backward-looking center ($Cb(U_{n-1})$ vs. $Cb(U_n)$). The original version of Centering (Grosz et al., 1995) distinguishes between three transitions with increasing disruptiveness: *Continue*, *Retain*, and *Shift*. Brennan et al. (1987), in their attempt to apply Centering to pronoun resolution, extend this scheme by differentiating the *Shift* transition into a *Smooth-Shift* and a *Rough-Shift* (Walker et al., 1994). Brennan et al. (1987) utilize the predictions made by the Centering algorithm with respect to the coherence of a discourse (expressed as one of four transitions) for pronoun resolution in the following way: When a pronoun is encountered in the current utterance, it is tentatively paired with all compatible referents from the previous utterance, forming a set of so-called *anchors*, i.e. possible bindings for the pronoun. Each of these bindings corresponds to a particular assignment of discourse entities to the *Cp* and the *Cb* of the current utterance. The resolution is then realized by selecting the binding which maximizes the local coherence, i.e. the one leading to the least disruptive transition. Brennan et al. (1987) demonstrate the functioning of their HPSG-based system with a small number of constructed example texts of three to four sentences in length. There is no large-scale quantitative evaluation. Kehler (1997) points out a couple of problems with Brennan et al. (1987)'s approach. One of these is the fact that it requires a whole utterance to be processed before a pronoun is resolved, although the original motivation of Centering included the claim that this was an implausible requirement for a theory of pronoun resolution because it failed to account for garden path phenomena, i.e. for the tendency of people to assign pronouns to referents even before the containing utterance is finished.

The pronoun resolution system RAP described in **Lappin & Leass (1994)** also belongs to the class of manually crafted systems, with the difference that it selects NP antecedents for pronouns based on their salience weights. These weights are calculated for NP antecedent candidates on the basis of syntactic and other properties, including e.g. grammatical role (subject vs. dir. object vs. indir. object/complement) or occurrence in an existential-*there* construction. In addition, the presence of parallelism (e.g. identical grammatical roles) between the current pronoun and an antecedent candidate can boost the salience of this candidate. In contrast to the algorithm by Hobbs (1978), the RAP system can be seen as an early example of the entity-mention approach. Although it considers individual antecedents only (like in the mention-pair approach), the salience of each antecedent is based not only on its own properties, but also on those of the pertaining discourse entity, i.e. the coreference class that the current antecedent candidate belongs to. The association between the antecedent candidate and the discourse entity is realized by assigning to the former the *sum* of the salience weights of all other NPs belonging to the same discourse entity. In order to prevent earlier NPs that are not referred to any more from accumulating salience weight, their salience weights are systematically decreased as the resolution proceeds.

In an evaluation on unseen data, the RAP system outperformed an implementation of the algorithm by Hobbs (1978) by 4%. Lappin & Leass (1994) conclude from this that, at least for English, their salience weights correspond strongly to the positional ranking yielded by the tree search algorithm of Hobbs (1978).

McCarthy & Lehnert (1995) describe the RESOLVE system, which is one of the first classification-based systems for coreference resolution. McCarthy & Lehnert (1995) employ a decision-tree learner (C 4.5 (Quinlan, 1993)) to build a classifier for automatically deciding whether or not a pair of noun phrases is coreferent. Thus, like most of the first classification-based systems, RESOLVE belongs to the class of mention-pair approaches which model coreference resolution as a binary classification task. In this paradigm, individual data instances consist of a tuple <ana, ante>, i.e. one anaphor and one antecedent. McCarthy & Lehnert (1995) create training and test data by exhaustively pairing every noun phrase in a document with every other noun phrase in the same document. The system uses only eight simple binary features, some of which are special to the domain of application, i.e. MUC-5. McCarthy & Lehnert (1995) compare the performance of RESOLVE to that of a manually crafted rule set using the same features. They find that the pruned version of RESOLVE has only a slightly worse precision[26] (92.4 vs. 94.4) but a significantly better recall (80.1 vs. 67.7) and thus also a better overall F-measure (85.8 vs. 78.9) than the rule set.

McCarthy & Lehnert (1995) already identify some of the major problematic issues of the binary classification approach and of coreference resolution evaluation. For example, they note that the combinatorial way of data generation leads to a strong bias of the resulting data set (and thus the learner) towards classifying mention-pairs as not coreferent. On the application level, this translates to a more conservative classifier that favours high precision over high recall. In the information extraction context in which McCarthy & Lehnert (1995) work, they regard this as a desirable property. In later work, however, unbalanced (or skewed) data sets came to be considered as a problem of machine-learning based coreference resolution (cf. Chapter 7.2.2). McCarthy & Lehnert (1995) are also among the first to draw attention to the fact that it is not sufficient to evaluate coreference resolution by counting how many anaphors were linked to their correct antecedent, because this disregards the *transitivity* of the relation. McCarthy & Lehnert (1995) instead calculate precision and recall on the transitive closures of all correct anaphoric links vs. all anaphoric links found by their system. In later work, more refined evaluation measures for coreference resolution were introduced (cf. Chapter 7.1).

Aone & Bennett (1995) is an approach that in many respects is similar to that of Mc-

[26] Calculated according to Vilain et al. (1995).

Carthy & Lehnert (1995). It also evaluates the performance of a classification-based machine learning system (C 4.5) by comparing its performance to that of a system based on manually designed knowledge sources. Unlike McCarthy & Lehnert (1995), Aone & Bennett (1995) work with Japanese text, and their feature set is much larger (66 partly domain-dependent features, as opposed to eight features used by McCarthy & Lehnert (1995)). The experiments by Aone & Bennett (1995) are more interesting since they systematically evaluate different strategies for data generation and training. One training data generation method (*anaphoric chain*), e.g. creates positive instances by pairing an anaphor with each preceding coreferent antecedent, while another considers the closest antecedent only. Both of these strategies are more constrained than the one used by McCarthy & Lehnert (1995) (cf. above). Aone & Bennett (1995) also contrast a strictly binary resolution (coreferent/not coreferent) with a multi-valued one in which the *type* of the relation is returned in case of coreference (*type identification*). Finally, Aone & Bennett (1995) employ different confidence thresholds for pruning. They found that all machine learning systems outperform the manual system, as long as the former use the *anaphoric chain* strategy. The best of these systems (using *anaphoric chain*, no *type identification*, and a confidence threshold of 75%) yielded a precision of 86.73, a recall of 69.73, and an F-measure of 77.30, as compared to 72.91, 66.51, and 69.57 for the manual system. Aone & Bennett (1995) note that, apart from the better performance, the machine learning system has the additional advantage of learning classifiers for types that the manually designed system did not even consider.

Baldwin (1997) describes the CogNIAC system for high-precision pronoun resolution. He justifies the focus on high precision – at the expense of recall – with reference to practical applications like Information Retrieval and Information Extraction. Baldwin (1997) argues that high precision is more desirable than high recall because precision errors are likely to cause more (or more serious) errors in the final application.

The system consists of an ordered set of manually designed rules that are implemented in Perl. Each rule specifies a particular constellation of anaphor and antecedent in which both expressions – if number- and gender-compatible – are very probably coreferent. CogNIAC thus belongs to the class of rule-based mention-pair systems. As an example, the **Unique Subject / Subject Pronoun** rule (Baldwin, 1997, p.40) states: If the subject of the prior sentence contains a single possible antecedent i, and the anaphor is the subject of the current sentence, then pick i as the antecedent. For each anaphor, rules are applied in a fixed order, and when a rule matches, the anaphor is resolved to the antecedent suggested by that rule. If no rule matches, or if two or more equivalent candidates are available, the anaphor is left unresolved. By this, Baldwin (1997) can ensure high precision of the resolution. It is also in this respect that CogNIAC differs from both rule-based and classification-based approaches like Hobbs (1978), Lappin & Leass (1994), McCarthy & Lehnert (1995), and Soon et al. (2001) which will always try to resolve an anaphor to some antecedent.

The system of **Soon et al. (2001)** can be regarded as a prototypical implementation of the classification-based mention-pair approach. Soon et al. (2001) use a C 5.0 decision tree learner to resolve coreferential relations between noun phrases in the MUC domains (here: MUC-6 and MUC-7). While McCarthy & Lehnert (1995) use a small set of eight features of which some are domain-dependent, Soon et al. (2001)'s set of twelve features is claimed to be domain-independent. It includes rather simple features like distance (in sentences) and number and gender agreement, but also semantic class agreement, where the semantic class is determined on the basis of WordNet (Fellbaum, 1998). Soon et al. (2001) also use a more sophisticated data generation algorithm than McCarthy & Lehnert (1995). For each anaphor, negative instances are created only for those antecedents occurring *before* the closest true antecedent. Also, only one positive instance is created for each anaphor, i.e. the one for the closest true antecedent. Another important difference in practical terms is that Soon et al. (2001) perform a fully automatic preprocessing. This also has a bearing on training and test data generation, since they can only consider those noun phrases that are detected and analysed by their preprocessing module.

The best performance of the system by Soon et al. (2001) (in precision, recall, and F-measure according to Vilain et al. (1995)) is 67.3, 58.6 and 62.6 for MUC-6 and 65.5, 56.10, and 60.4 for MUC-7. Soon et al. (2001) also perform a couple of baseline experiments in which they run their classifier with an extremely reduced feature set (often a single feature only). It is revealing to see the surprisingly good performance of some of these systems. For example, if only the binary string match feature[27] is used, the system performance (precision, recall, and F-measure) is 65.6, 45.7 and 53.9 for MUC-6 and 71.4, 43.8 and 54.3 for MUC-7. This is clearly no evidence for good performance of the features, but rather an artifact of the MUC data sets. Even putting all noun phrases into

[27] Defined as string identity after determiners (if any) have been removed.

the same coreference class still yields a performance of 31.8, 89.9 and 47.0 for MUC-6 and 30.5, 87.5 and 45.2 for MUC-7.[28]

The approach of **Luo et al. (2004)** belongs to the class of classification-based approaches that overcome the limitation of the mention-pair approach, instead realizing the entity-mention approach. In Luo et al. (2004), the search space for coreference resolution is represented by means of the Bell tree. The Bell tree is a tree representation of all possible ways in which a given number of elements (i.e. mentions) can be partitioned into non-empty, disjoint subsets (each corresponding to an entity). In the tree, each leaf node corresponds to a unique partitioning of all mentions, and coreference resolution is thus modelled as traversing the tree in order to find the leaf corresponding to the correct partitioning. This traversal is implemented by moving through the list of all mentions in discourse order. For each mention, it has to be decided whether it starts a new entity or whether it has to be linked to one of the existing ones. The approach by Luo et al. (2004) is global and incremental in that it does not consider candidate *mentions*, but candidate *entities*, which are represented as the set of all mentions of a referent. This way, earlier resolution decisions can be taken into account when making the current decision. The actual classification is done by a maximum entropy classifier which is trained to compute the linking probability between an entity and a mention. This classifier uses a fairly standard set of features, comparable to that employed by Soon et al. (2001). The probabilities returned by the classifier are used to implement a heuristic search method for the Bell tree. This is necessary because an exhaustive search and optimization is intractable due to the exponential growth of the tree. This dependence on a heuristic is one of the points of criticism brought up against the approach by Luo et al. (2004), e.g. by Nicolae & Nicolae (2006).

Yang (2005) describes a coreference resolution approach that is also classification-based, using the C5.0 decision tree learner. His so-called Twin-Candidate Model is closer to the mention-pair approach than to the entity-mention approach in that it does not use representations of entities. It does, however, to some degree overcome the local limitation of the common mention-pair approach by learning antecedent preferences. Unlike in the mention-pair approach, data instances in the Twin-Candidate Model consist of one anaphor and two antecedent candidates. An instance is thus formed by a triple

[28] The relatively low recall of this baseline is caused by faulty preprocessing which fails to detect all noun phrases in the input.

<ana, ante1, ante2>. For each instance, either ante1 or ante2 (but not both) is always a true antecedent of the anaphor. The information which of the two antecedent candidates is the true one is encoded in the binary class label of each instance: If ante1 is the true antecedent, the instance is labelled as *10*, if ante2 is the true antecedent, the instance is labelled as *01*. Each instance contains features (in the form of attribute-value pairs) describing the following aspects of the instance: Features of ana, features of ante1 and ante2, features of the relation between ana and ante1 resp. ana and ante2, and features of the relation between ante1 and ante2. The information encoded by these features is comparable to that used in other feature sets, e.g. by Soon et al. (2001). The features describing the relation between ante1 and ante2 encode potential preferences of one antecedent candidate over the other. Examples include **inter_BetterStrSim** and **inter_BetterSemSim**. Both are three-valued features that return the number of the antecedent (1 or 2) which has the highest score in the string similarity resp. semantic similarity[29] feature, or 0 if both candidates score equally.

Training instances in the Twin-Candidate Model are created by finding for each anaphoric expression the immediate antecedent, and by combining this as the true antecedent with all non-coreferent antecedent candidates appearing before the anaphor. This way of training data generation avoids the data skewness problems associated with the common binary classification paradigm, because there are no more negative instances. Instead, in the training data created on the basis of the MUC-6 data set, there are 16329 *01* instances and 8920 *10* instances. During testing, a given anaphor is combined with all potential candidate pairs and submitted to the classifier. For each instance thus formed, the classifier returns which of the two candidates is the more probable antecedent for the anaphor. After all instances have been classified, the candidate that has been proposed as the correct one most often is selected as the most probable antecedent.

One interesting feature of Yang (2005)'s approach is that it provides an elegant mechanism to integrate coreference resolution with anaphoricity determination. The idea is to also include into training those noun phrases as potential anaphors that are actually non-anaphoric, and to distinguish the resulting instances using a special (third) class label. If a non-anaphoric noun phrase is encountered during training, instances are created for it by pairing it with pairs of antecedents just like for normal instances. The difference is that since none of the two candidates is the true antecedent, the resulting instances are labelled as *00*. This way, the classifier is able to learn that for certain types

[29] Semantic similarity is calculated on the basis of WordNet.

of anaphors no preference should be given to any of the two candidates. Due to the rather unconstrained way in which training instances are created, the *00* class dominates the training data. In the training data created on the basis of the MUC-6 data set, e.g., 62097 instances (68.5% of all instances) are of this class.

Yang (2005) reports performance figures (precision, recall, and F-measure according to Vilain et al. (1995)) of 71.3, 65.8 and 68.4 on MUC-6 and 68.9, 65.2 and 67.0 on MUC-7. In particular precision and F-measure are competitive, which Yang (2005) mainly attributes to the effect of the integrated anaphoricity determination.

Nicolae & Nicolae (2006) in some sense reverse the coreference resolution process. They start out not with a set of individual mentions that need to be combined into coreference sets, but with a small number of coreference sets each of which initially contains *all* mentions of a particular ACE entity type (PERSON, ORGANIZATION, etc.). The task is then to partition each of these sets into subsets of actually coreferring mentions. The initial partitioning according to entity types is justified by the assumption that mentions of different types cannot be coreferent. Nicolae & Nicolae (2006) adopt a graph representation for coreference sets in which each mention is a node that is linked to all other nodes in the set. Each edge between two nodes is assigned a numerical value representing the confidence that these two nodes are coreferent. The values are obtained from a statistical pairwise coreference model, based on a modified variant of the maximum entropy model used by Luo et al. (2004). The actual coreference resolution is performed by repeatedly cutting the graphs into subgraphs until a certain stopping condition is fulfilled. From all possible cuts that could be performed on a given graph, the one is selected that minimizes what Nicolae & Nicolae (2006) call the cut weight. The cut weight is defined as the number of mentions that are correctly placed in their set. This number is estimated on the basis of the confidence values assigned to the edges connecting each graph (i.e. each set). The stopping condition which controls whether or not a graph should be cut is provided by a separate machine learning classifier. This binary classifier was trained on pairs of graphs (each pair resulting from a particular cut) to decide whether this cut makes the result better or worse.

The main advantage of the approach by Nicolae & Nicolae (2006) is that it allows for globally optimized clustering. As such, it is superior to approaches which use only local information to decide which of several potential antecedents an anaphor should be linked to. This, however, is true only for lexical mentions and not for pronouns:

As Nicolae & Nicolae (2006) point out, they entirely exclude pronouns from the graph cutting procedure because these were found to be linking too liberally to many different potential antecedents. Instead, pronouns are resolved only after the graph cutting procedure has terminated, by simply linking them to the antecedent with the highest pairwise coreference confidence. Thus, when it comes to pronouns, the approach by Nicolae & Nicolae (2006) is equivalent to the mention-pair approach.

Bergsma & Lin (2006) present a corpus-based approach for improving standard classification-based mention-pair systems for pronouns with a feature they call *path coreference*. They start with the observation that the correct resolution of cases like

[John]$_i$ needs [his]$_i$ friend.

and

[John]$_i$ needs [his]$_j$ support.

requires a type of world-knowledge that is not available in any current pronoun resolution system. Rather than manually creating rules that state that *noun* and *pronoun* in a construction of the form "*Noun* needs *pronoun*'s support" are most likely not coreferent, Bergsma & Lin (2006) propose to acquire these regularities automatically from a huge (85 GB) dependency-parsed corpus of news articles. A key concept in their approach is that of gender and number compatibility. Bergsma & Lin (2006) observe that dependency paths involving incompatible (and thus non-coreferent) noun-pronoun pairs like "John needs her support" or "They need his support" are much more frequent in their corpus than the same path involving compatible noun-pronoun pairs, like "John needs his support". They conclude from this that noun and pronoun in the path "*Noun* needs *pronoun*'s support" are most probably not coreferent. Based on a similar argumentation, they add the path "*Noun* needs *pronoun*'s friend" to the list of patterns that signal that noun and pronoun are likely (but not necessarily) coreferent. With this method, they mine several million dependency paths from their corpus, each of which is associated with a confidence value based on the corpus counts. Bergsma & Lin (2006) describe several ways to put this information to use for pronoun resolution, one of which is as a simple boolean feature stating whether or not path coreference has been determined for a given pair of anaphor and antecedent.

To sum up this chapter, the following observations concerning pronoun resp. coreference resolution in written text are worth noting. Early approaches like Hobbs (1978), Brennan et al. (1987), McCarthy & Lehnert (1995) and Baldwin (1997) are limited in that they consider the resolution as a task to be solved locally, i.e. by only taking information of anaphor and potential antecedent into account (mention-pair). An exception is Lappin & Leass (1994), who already incorporate some incrementality into the resolution process by allowing the salience of an antecedent noun phrase to be increased by the salience of its antecedents. Most of these early approaches are symbolic or rule-based. McCarthy & Lehnert (1995) is one of the first approaches to introduce the machine-learning based binary classification paradigm. This paradigm, realized in an exemplary fashion by Soon et al. (2001), has been predominant until a couple of years ago. Luo et al. (2004) and Nicolae & Nicolae (2006) are more recent classification-based systems that attempt to overcome the limitation of mention-pair approaches. They optimize the process of resolving an anaphor by considering all coreference sets created up to the point of the anaphor (Luo et al., 2004) respectively all sets globally (Nicolae & Nicolae, 2006). Bergsma & Lin (2006), finally, is an original approach that is different from the others described so far in that it targets a particular sub-problem of coreference resolution only, employing probabilistic counts from a huge unannotated corpus.

Another finding of this chapter relates to the role of linguistic knowledge resp. linguistically motivated features for pronoun resp. coreference resolution in written text. A common linguistically motivated constraint on coreference is **number and gender agreement**. It can be found in virtually all approaches, beginning with Hobbs (1978), where it is in fact the only constraint on an otherwise unconstrained but highly linguistically motivated tree-traversal antecedent search, up to the entirely different and purely statistically based approach of Bergsma & Lin (2006). A second linguistic concept is that of **(local) coherence**. It is employed e.g. in Centering-based approaches like that of Brennan et al. (1987), where it is operationalized as grammatical parallelism (anaphor and antecedent are preferred to have the same grammatical function) and as subject preference (the antecedent is preferrably the grammatical subject of the previous utterance). The later preference is encoded in the ranking order of the list of forward-looking centers, which identifies the subject as the highest ranked discourse entity. A third linguistic concept that is the basis of the approach by Lappin & Leass (1994) is that of **salience**. Partly, this concept is operationalized in terms of the same grammatical phenomena as coherence, i.e. subject preference and grammatical parallelism, but

Lappin & Leass (1994) also include syntactic devices like existential *There*-constructions. A final important point of this chapter is that many systems use the MUC or ACE data sets for training and testing. These data sets have originally been created to foster research on pronoun resp. coreference resolution as a preprocessing step in practical applications (like Information Extraction or Information Retrieval). While coreference resolution with time has been decoupled from these 'downstream' tasks, some (especially earlier) works explicitly draw some of their motivation from the potential contribution of coreference resolution in practically usable systems (e.g. McCarthy & Lehnert (1995) and Baldwin (1997)).

In the next chapter, we will now turn to a description of the state of the art in spoken dialog pronoun resolution. As mentioned in the introduction to this chapter, a distinction will be made between unimplemented algorithms and implemented systems. It will become clear that while there is some theoretical work on the first aspect, practically usable automatic resolution systems are not even remotely as far developed as they are for written text.

5.2 Unimplemented Algorithms For Spoken Dialog Pronoun Resolution

5.2.1 Byron & Stent 1998

Byron & Stent (1998) describe a first attempt to adapt Centering to dialog. Their main motivation is that Centering addresses both generation and understanding of referring expressions, and that it might therefore be an appropriate theoretical basis for practical spoken dialog systems. Although Byron & Stent (1998) mention the application of Centering to pronoun resolution only as future work, their work is important here because the applicability of Centering to spoken dialog is obviously a prerequisite for Centering-based pronoun resolution (along the lines of Brennan et al. (1987), Chapter 5.1). Their starting point is the observation that Centering uses a conceptual machinery that makes assumptions that are met by written, monological text, but not by two-party dialog. The four most important open issues mentioned by Byron & Stent (1998) are the following. First, since antecedent candidates are proposed on the basis of the entities in the *Cf* list, the delimitation of the dialog into utterances is a crucial factor. Utterance segmentation in spoken dialog, however, is much more difficult than in written text. Second, the grammatical subject of the current utterance occupies a prominent position in the pertaining *Cf* list, the *preferred center* (*Cp*). In spoken dialog, the subjects are often the speaker resp. the hearer, represented as 1st resp. 2nd person pronouns, and it is unclear whether these should be included in the *Cf* list at all. Third, the Centering algorithm depends on a linear ordering of utterances, so that the previous utterance can be determined. In two-party dialog, where each utterance is associated with an individual speaker, it is unclear whether *previous* should mean *immediately preceeding*, regardless of speaker, or rather *most recent utterance by the same speaker*. Finally, utterances in spoken dialog may be (partially or completely) abandoned, or they may not contain any discourse entities for other reasons. It is unclear how these utterances should be handled.

Byron & Stent (1998) manually evaluated three versions (models) of Centering that differed from the original algorithm (Grosz et al., 1995) and from each other in how they handled some of the issues described above. Testing was performed on a corpus of four dialog transcripts from the CALLHOME corpus, consisting of 664 non-empty utterances. Prior to the application of the models, utterance segmentation was done jointly by the authors. Then, each author manually applied each of the three models to three transcripts.[30] From the resulting annotations, Byron & Stent (1998) created a consolidated version, which served as the basis for the evaluation of the three models. Three evaluation criteria were used: The percentage of cases in which a model leaves an empty *Cb*, the percentage of cases in which a model predicts the real topic of an utterance (as determined by the annotators) as the *Cb*, and the percentage of cases in which a model predicts a *cheap* or *expensive* transition between utterances (Strube & Hahn, 1996). The model that fared best in all three criteria was Byron & Stent's model 1. This model did allow 1st and 2nd person pronouns to appear in the *Cf* list and to function as *Cb* as appropriate, and it considered the immediately preceeding utterance as the previous one, regardless of which speaker it was associated with. In absolute terms, however, the performance of even the best model for detecting the *Cb* of spoken dialog utterances (which is a prerequisite to using it for pronoun resolution) leaves a lot to be desired: Only for roughly half of the utterances, a *Cb* was found at all. Of these, 57% were realized by 1st or 2nd person pronouns, which are trivial to resolve. For utterances where a *Cb* was found, it was correct in only 35% of cases. The percentage of cheap transitions, which corresponds to the human notion of coherence, was only 41% for model 1. Byron & Stent (1998) come to the conclusion that although some basic assumptions of Centering theory hold in dialog as well, the performance of even their best model is too low to be

[30] One transcript was used for training the annotators.

practically useful.

5.2.2 Rocha 1999

The work by Rocha (1999) (based on Rocha (1997)) is similar to that by Schiffman (1985) (described in Chapter 2.3) in that both authors attempt to build statistical models of coreference in spoken dialog. In contrast to Schiffman (1985), Rocha (1999) uses a more heterogeneous corpus consisting of partial dialogs from the London-Lund corpus. Also, he includes all types of pronouns, except zero pronouns, and also adverbials and nominals. His corpus contains 3090 cases, approx. 50% of which are pronouns. Rocha (1999) also includes antecedents that are discourse chunks (i.e. what Webber (1991) calls *discourse deictic* and what in this thesis is treated as one of several types of *vague*.) In a first annotation phase, Rocha (1999) identifies topical discourse entities, on the grounds that topicality is assumed to play a crucial role for coreference. Topical discourse entities (or simply *topics*) are identified on different levels. Rocha (1999) assumes a single global **discourse topic** for the entire dialog, but concedes that a dialog can be split in case "there is a radical and stable change of topic within the dialogue" (Rocha, 1997, p. 55). Topics are also identified on the level of segments and subsegments. Rocha (1999) also annotates discourse entities as (local or global) thematic elements, i.e. discourse entities that are *related* to topics. Rocha (1999) manually annotates anaphor-antecedent pairs in his corpus, identifying the following four major descriptive properties. The property **type of anaphor** classifies the anaphoric word or phrase with one of 27 categories, including **subject pronoun**, **noun phrase**, or **One-anaphor**. The property **type of antecedent** identifies the referent of the anaphor, and also specifies whether the referent is explicitly or implicitly mentioned. For non-referential pronouns, Rocha (1999) uses a special value. In his corpus, 2562 cases (82.91%) are classified as explicit, 412 (13.33%) as implicit, and 116 cases (3.3%) are classified as non-referential (Rocha, 1997). Based on the previous topic annotation, the property **topicality status of antecedent** states whether the current antecedent plays a local, global, or sublocal topical role. For anaphors whose antecedents are altogether missing (i.e. non-referential anaphors), or those that are too vague to be assigned a topical role, the special value **focusing device** is used. Finally, the property **processing strategy** distinguishes anaphor-antecedent pairs with respect to the type of knowledge that is required for identifying them as coreferent. Rocha (1999) identifies four types of processes:

- **Lexical processes.** Lexical repetition, lexical relations like part-whole. (35.4%)

- **Discourse processes.** Full processing of combined bits of discourse information. (16.3%)

- **Collocations.** Knowledge about collocational constructions like *I mean it* or *That is to say*. (9.0%)

- **Syntactic processes.** Heuristics like first-candidate, or syntactic parallelism. (39.3%)

Given the data annotated according to the above categories, Rocha (1999) develops what he calls the *antecedent-likelihood* theory. It is based on a probabilistic model of associations and conditional probabilities derived from frequencies observed in the annotated data. Rocha (1999) states that this model can be used to control the flow of processing in spoken dialog anaphor resolution. He postulates the category **type of anaphor** as the starting point of the resolution process in this model, mainly on the grounds that the value for this category is easy to determine automatically. Once this value is known, the distribution found in the annotated data is used to determine which processing strategy is the most promising for the given type of anaphor. The choice of the category **processing strategy** as the second category to be checked after **type of anaphor** is grounded on a statistical association test on the data that showed that "once the distribution of type of anaphor is known, the chances of predicting the processing strategy correctly are forty-one percent higher" (Rocha, 1997, p. 57). Since the model only produces an optimized sequence for candidate checking, other measures are required to check whether the candidate returned by the currently checked processing strategy is correct. For this, Rocha (1999) suggests to use semantic selectional restrictions and what he calls an *association history*. This latter device keeps track of the verbs that antecedent candidates were associated with (i.e. their predicative contexts), and uses this information to boost or reject a candidate based on its current predicative context.[31]

5.2.3 Eckert & Strube 2000

One of the earliest empirically based works that explicitly adresses pronoun resolution in spoken dialog is Eckert & Strube (2000). The data collection and annotation pertaining to this work has already been described in Chapter 4.2.1. As was mentioned there, Eckert & Strube (2000) manually enrich their corpus (a part of the Switchboard corpus of two-party telephone conversations, cf. Chapter 3.2) with an extensive and detailed

[31] See our **FullVerbIdent** feature in Chapter 6.2.2.

dialog act annotation. Dialog acts are further grouped into so-called *synchronizing units* (SUs). Dialog acts and SUs provide the major units on which the resolution algorithms operate. In the present chapter, the focus is on these algorithms and on the knowledge sources that they employ.

Eckert & Strube (2000) outline two separate algorithms for identifying the antecedents of personal and demonstrative pronouns. The algorithms are rather abstract, mainly specifying the order in which certain tests are to be performed and tentative resolutions are to be attempted. The tests make reference to properties of the anaphor that Eckert & Strube (2000) call **I-Incompatible** and **A-Incompatible**. These properties are intended to capture the incompatibility of the anaphor with an individual (NP) or abstract (VP) antecedent. Eckert & Strube (2000) give some examples of when these properties hold for an anaphor, but do not provide an operationalization that could be used for automatically identifying them.[32] Eckert & Strube (2000) use two resolution functions, one for individual anaphors (*resolveInd*) and one for discourse-deictic ones (*resolveDD*). Both functions basically only implement a linear search through a candidate list. For *resolveInd*, this list is the S-list (Strube, 1998), i.e. the list of all discourse entities referred to by NPs (regardless of whether they have been anaphorically rementioned before). For *resolveDD*, it is the so-called A-list, which contains only those discourse entities that were referred to anaphorically before. Both lists are emptied every time an SU has been completed. This way, the antecedent search space is kept small. If an A-Incompatible anaphor is encountered, the S-list is searched for a matching (i.e. number and gender compatible) antecedent. If an I-Incompatible – and thus potentially discourse-deictic – anaphor is encountered, the A-list is searched first for a VP referent that has recently been referred to anaphorically. This is necessary to enable the algorithm to resolve multiple rementions of the same VP referent, i.e. anaphoric *chains* with VP referents. Only if none is found, the potentially discourse-deictic anaphor creates (coerces) a new VP referent from the preceeding discourse, using a mechanism called Context Ranking. This mechanism searches the previous context for a complete clause to be used as the VP antecedent. The search is controlled by the linear order of dialog acts, and also takes into account if dialog acts are incomplete or abandoned. If this is the case, they are treated as invisible to the algorithm, and simply ignored. The Context Ranking algorithm is also inspired by the notion of the *Right Frontier* (Webber, 1991). Webber (1991) describes the accessibility of discourse segments in terms of their position in a discourse structure

[32] An attempt to operationalize some of these properties for automatic identification is described in Chapter 6.2.1.

representation that she assumes to be tree-like. In this representation, only nodes on the rightmost branches (the *right frontier* of the tree) are assumed to be accessible for discourse-deictic reference. Eckert & Strube (2000) note that this can also be interpreted in terms of (linear or hierarchical) adjacency. They utilize it in their algorithm by considering as potential antecedents for discourse-deictic pronouns only those clauses that are the *rightmost* ones in their respective dialog acts.

The algorithms for resolving personal resp. demonstrative pronouns are identical except for their treatment of expressions that do not exhibit I- resp. A-Incompatibility. In these cases, Eckert & Strube (2000) implement a preference for demonstratives to be discourse-deictic by first submitting them to the *resolveDD* function. Only in case of failure of the *resolveDD* function will a demonstrative be submitted to the *resolveInd* function. For personal pronouns, Eckert & Strube (2000) propose to use *resolveInd* first and then *resolveDD*. If none of the resolution function succeeds, both demonstratives and pronouns are classified as *vague*.

Eckert & Strube (2000) report results of the manual application of the algorithms to a set of three dialogs (199 expressions, including other pronouns than *it*, *this*, and *that*). They manually excluded pleonastic and discarded expressions, thus preventing them from triggering resolution attempts. Eckert & Strube (2000) report a precision and recall of 66.2 resp. 68.2 for pronouns and 63.6 resp. 70.0 for demonstratives.

5.2.4 Navarretta 2004

Navarretta (2004) describes an algorithm for the resolution of (in her terminology) individual and abstract pronominal anaphors (IPAs resp. APAs) in Danish dialog. Details about her corpus (e.g. number of pronouns) are not given, but Navarretta (2004) reports that the κ agreement for the classification of pronouns as IPA vs. APA was $.86$. The algorithm is strongly influenced by the work of Eckert & Strube (2000). For APAs, Navarretta (2004) uses the same resolution strategy than Eckert & Strube (2000), including the Context Ranking algorithm. For IPAs, Navarretta (2004) proposes to incorporate the notions of salience and givenness into the resolution. At the same time, she makes her algorithm sensitive to speakers' explicitly changing the salience of referents by means of linguistic devices related to the information structure of the utterance. Among the devices she considers are e.g. cleft or existential constructions, focussing adverbs, or prosodic markedness. She also proposes to explicitly take question-answer-pairs into account. Another aspect in which the algorithm by Navarretta (2004) extends that by

Eckert & Strube (2000) is that she gives an explicit ranking of preferences for individual and abstract antecedent selection. In this ranking, *syntactic parallelism* receives the highest preference, which means that information structural aspects of antecedents' salience can be overriden by other antecedents on syntactic grounds. The second main contribution of Navarretta (2004) consists in the adaptation of the algorithm by Eckert & Strube (2000) to the Danish language. In doing so, she also introduces some language-specific features, including special resolution strategies for particular Danish pronouns, based on preferences derived from corpus inspection.

Navarretta (2004) reports results for the manual application of her algorithm of 74.87 precision resp. 68.81 recall for IPAs, and 71.84 precision resp. 73.98 recall for APAs. Information about the test dialogs (except that they have been chosen at random) is not provided.

5.3 Implemented Spoken Dialog Pronoun Resolution Systems

5.3.1 Strube & Müller 2003

To our knowledge, Strube & Müller (2003) is the only implemented system so far that resolves normal and discourse-deictic pronouns in unrestricted spoken dialog. The system uses a decision-tree learner (CART (Breiman et al., 1984)) and runs on 20 dialogs from the Switchboard portion of the Penn Treebank (Marcus et al., 1993). Strube & Müller (2003) is a straightforward application to spoken dialog of the mention-pair, binary classification paradigm known from pronoun resolution in written text. The major extension it contributes is the definition of a couple of features that are specifically tailored for the detection of antecedents of discourse-deictic anaphors. One group of these new features capture a pronoun's preference for (in our terminology) NP or VP antecedents by taking the pronoun's governing verb into account. The preferences are encoded as the relative frequencies with which a verb subcategorizes a noun, a verb, or a sentence complement. Rather than using an available list of subcategorization preferences (like e.g. that of Briscoe & Carroll (1997)), Strube & Müller (2003) chose to compile their own list on the basis of 553 syntactically parsed Switchboard dialogs. Another group of features captures the relative importance of non-NP antecedents in terms of TF*IDF and Information Content (IC) (Baeza-Yates & Ribeiro-Neto, 1999). The annotation by Strube & Müller (2003) utilizes the manually created VP and S constituents from the Penn Treebank for the identification of non-NP antecedents. Therefore, these antecedents consist of *sequences* of words, and Strube & Müller (2003) characterize each one by means of its average TF*IDF resp. average IC score. Individual TF*IDF and IC scores are calculated on the basis of all 553 dialogs in the corpus. Another aspect in which Strube & Müller (2003) extend the normal binary classification paradigm is by providing a mechanism for the creation of candidate antecedents for discourse-deictic pronouns. If the current anaphor is *it* or *that*, Strube & Müller (2003) create non-NP antecedent candidates by extracting all S and VP constituents from the last two valid[33] sentences. From these, they filter out inaccessible constituents, using an accessibility criterion which is a shallow approximation of the *right frontier* condition (Webber, 1991). The remaining constituents are paired with the current anaphor to form training and testing instances.

For *it*, *this* and *that*, Strube & Müller (2003) report a best performance for dialog-wise, i.e. 20-fold cross validation of 40.41 precision and 12.64 recall, calculated according to Vilain et al. (1995). The recall is not representative because it is calculated against all correct coreferential links in the corpus, and not just those with pronoun anaphors. Strube & Müller (2003) do not give separate performance figures for NP and non-NP antecedents, so that nothing can be said about the performance of the resolution of discourse-deictic pronouns. In their feature evaluation, Strube & Müller (2003) note that the features encoding subcategorization preferences were mostly ignored by their decision tree learner and thus did not contribute to the performance of their system.

While the system of Strube & Müller (2003) is fully implemented, it is not entirely automatic, since it draws a lot of non-trivial information from the Penn Treebank. First, markables (including information about grammatical function etc.) are derived directly from the syntactic constituents available in the treebank. These markables are the basis for both the coreference annotation – which was performed by the authors – and the training and test data generation. The grammatical function information is employed in the form of features describing the grammatical function of anaphor and antecedent, and for establishing what Strube & Müller (2003) call *syntactic parallelism*, i.e. the identity of grammatical function of anaphor and antecedent. Second, based on the `-UNF`-tag, markables ocurring in unfinished utterances (i.e. disfluencies) are singled out. Third, the hierarchical structure of the syntactic constituents is utilized to implement the accessibility filter for potential candidates for discourse-deictic pronouns.

[33] In Strube & Müller (2003)'s terminology, a *valid* sentence is one which is neither a backchannel nor unfinished.

5.3.2 Byron 2004

Byron (2004) describes an implemented system for resolving personal and demonstrative pronouns in task-oriented TRAINS dialogs (cf. Chapter 3.2). The system, which is called PHORA, runs on the data set whose creation was described in Chapter 4.2.1. PHORA is considerably different from the other implemented system by Strube & Müller (2003). It is a rule-based, non-probabilistic system which is built upon a domain model which contains explicit representations for the semantic types of concrete and abstract objects (including situations, actions, and events) relevant for the TRAINS domain (cf. Chapter 4.2.1). Examples of the former are *Boxcar*, *Tanker*, or *Orange Juice*, while examples of the latter are *Load*, *Depart*, *TakeTime*, *Arrive*, or *Unload*. The types are organized hierarchically, such that *Orange Juice* is represented as a *Liquid Commodity* which is a kind of *Cargo* which is a kind of *Movable-object*. Predicates in the lexicon of the system's parser are associated with semantic restrictions which are specified in terms of the semantic types in the domain model. One restriction e.g. states that for something to be the subject of the verb *happen*, it must be of type *Event*. More domain-specific constraints include one that states that something can only be attached to an engine if it is of type *Container*.

Another information source for Byron's system comprises manual feature annotations that were performed alongside the coreference annotation described in Chapter 4.2.1. The features fall into two broad categories: features of the pronoun and features of the linguistic antecedent. Features to be annotated for each pronoun include:

- **Clause level of the Pronoun.** Whether the pronoun occurs in a main or subordinate clause.
- **Grammatical Role.** Whether or not the pronoun is the subject of its clause.
- **'It' in same utterance.** Applicable to demonstratives only: Whether or not an instance of *it*, *its*, *they*, *them*, or *their* occurs in the same utterance. (dropped)
- ***Do* + pronoun construction.** Whether or not the pronoun is the object of *do*. (dropped)

In addition, the following features were annotated for the linguistic antecedent:

- **Form.** If a linguistic antecedent is available, whether it is a noun phrase, pronoun, or non-noun phrase, else none.

- **Clause level of Linguistic Antecedent.** Whether the linguistic antecedent occurs in a main or subordinate clause.
- **Grammatical Role of Linguistic Antecedent.** Whether or not the linguistic antecedent is the subject of its clause.
- **Distance between Linguistic Antecedent and Pronoun.** Whether the linguistic antecedent and the pronoun occur in the same, adjacent, or remote utterances.
- **Linguistic Antecedent Relation to Semantic Antecedent.** Whether the relation between the semantic and the linguistic antecedent is one of coreference or some other type. (dropped)

In the first phase of pronoun resolution, the domain model and semantic restrictions are employed to compute the semantic type of the pronoun's referent based on the pronoun's predicative context. Depending on the required semantic type of the referent and the form of the anaphor (personal or demonstrative pronoun), in the second phase either an existing referent is selected, or a new one is created. In either case, access is made to an explicit discourse model, which is another integral part of PHORA. The discourse model is updated after each main clause resp. turn, to reflect the currently available referents for individual and discourse-deictic anaphors. The former are represented in the discourse model as DEs (discourse entities), while the latter are merely DE *proxies*. This distinction is necessary because Byron (2004) wants to identify not only the surface antecedent (i.e. the VP or sentence) for discourse-deictic anaphors, but also the semantic type of the corresponding referent. Since the same VP or sentence can give rise to referents of different semantic types, and since it is inefficient to create them all in advance, Byron (2004) uses proxy representations. An actual referent is not created from a proxy until an anaphor is encountered whose predicative context (cf. above) identifies it as requiring a referent of a particular semantic type (e.g. an event or a proposition). The actual referent creation is modelled by Byron (2004) in terms of a set of *referring functions* like **Event(p)** or **Prop(p)**. These functions take a proxy as argument and return a referent of the particular semantic type.

In order for the semantic machinery described above to work, Byron (2004) depends on highly accurate preprocessing that will produce correct syntactic and semantic analyses of the dialog utterances. Since for raw dialog transcripts (even from the rather simple TRAINS domain) the required degree of accuracy cannot be provided fully automatically, a number of measures are employed by Byron (2004). First, for evaluation she uses a corpus that underwent a manual cleanup procedure. In the course of this procedure, speech disfluencies and discourse markers were removed in order to make the utterances parsable. Sometimes, also the wording of utterances was slightly edited or simplified, while the original utterance meanings were maintained. Another simplification that Byron (2004) makes is that she manually corrects errors from 'upstream' preprocessing components (speech recognizer, parser, semantic analyser), thus preventing these errors from impairing the pronoun resolution.

Byron (2004) does an evaluation of PHORA on a small set of (cleaned-up) TRAINS dialogs. The system achieves a precision of 75.0 and a recall of 65.0 for *it* (50 instances) and a precision of 67.0 and a recall of 62.0 for *that* (93 instances) if all available semantic restrictions are used. Precision drops to 52.0 for *it* and 43.0 for *that* when only domain-independent restrictions are used. Just like Strube & Müller (2003), Byron (2004) does not provide separate figures for the performance of anaphors with NP and non-NP antecedents. She notes, however, that the inclusion of non-NP antecedents into her algorithm brings about an increase in precision from 54.0 to 74.0 for *it* and from 11.0 to 54.0 for *that*. The much higher relative gain for *that* is probably due to the fact that *that* is more often discourse-deictic than *it* and that it is thus more dependent on the availability of non-NP antecedents.

5.4 Chapter Summary

This chapter described some previous work on pronoun resp. coreference resolution in written text and spoken dialog. For the domain of written text, the current state of the art can be summarized as follows: There is a considerable number of implemented systems with a degree of robustness that makes them usable in real-world applications. Many of these systems employ the so-called mention-pair approach in which pronoun resp. coreference resolution is modelled as the mapping of anaphors to individual antecedents. The system by Soon et al. (2001) is representative of this class. More recent approaches attempt to overcome the limitations of the mention-pair approach by taking a more global perspective on the task. In general, the level of sophistication tends to be high, which can be seen in the fact that rather powerful and complex computational and statistical methods (Bell tree, graph-based methods) are applied. It is also obvious in the fact that there are implemented approaches which target specific subproblems resp. which apply rather specific techniques to improve the resolution of certain coreference phenomena. There are also numerous features employed in resolution that are based on linguistic phenomena.

For spoken dialog, on the other hand, the situation is different. Most of the few works on spoken dialog pronoun resolution are more descriptive and theoretically inclined, rather than aimed at building a practically usable system. One of the reasons for this is that pronoun resolution for spoken dialog does not have as obvious practical applications as written text. Another reason is that it has requirements – like disfluency detection, parsing, and discourse structure analysis (for discourse deixis) – which are beyond the current state of language processing. The following linguistically motivated features can be found in approaches for pronoun resolution in spoken dialog. On the one hand, for NP antecedents we find the same features that are also employed for written text, including e.g. number and gender agreement and grammatical parallelism. More interesting, however, are linguistic phenomena that are particular to the task of finding VP antecedents. Among these are e.g. **discourse structure** in terms of discourse topics (Rocha, 1999) resp. dialog acts (Eckert & Strube, 2000). A related notion is that of the ***right frontier***, which uses a tree-like discourse structure to define accessibility constraints for VP antecedents. Finally, a linguistically motivated phenomenon which is employed to control resolution algorithm flow is the **preference of demonstratives to be discourse-deictic** (Eckert & Strube, 2000; Byron, 2004). It is apparent that most of the linguistically motivated features relevant for spoken dialog are very demanding and clearly beyond the capabilities of current NLP technology. As a result of all this, there is only a very small number of implemented systems, and even these either implicitly or explicitly make assumptions that prevent their application in a real-world setting. This thesis, and in particular the work described in the next chapter, is thus a first attempt to fill this gap.

6 Practical Pronoun Resolution in DIANA-Summ

This chapter deals with two major practical aspects of our pronoun resolution system. The first aspect, described in Chapter 6.1, relates to the preprocessing that was performed in order to turn the raw textual data of the ICSI Meeting Corpus and the manual annotations into a cleaner, richer and more structured format. As was already mentioned, the main methodological requirement of the DIANA-Summ project was to build a dialog pronoun resolution system that is practically usable in a realistic application setting. As a consequence, any form of *manual* preprocessing like disfluency removal or filtering of non-referential pronouns was ruled out.[34] Instead, the entire preprocessing should be done fully automatically. This is a major point of difference between the work described in this thesis and the other implemented systems described in Chapter 5.3. The only concession in this respect is the decision to make full use of the information available in the manually transcribed ICSI Meeting Corpus, even though this information is much more detailed and accurate than what can be expected from current automatic speech recognition systems. Using this information, however, is justified because the focus of this thesis is on pronoun resolution, and the assumption of correct textual representation is only a minimal one.

Chapter 6.2 describes the second practical aspect of our system, the feature representation. This is required for modelling the output of the preprocessing in such a way that it is amenable to classification methods based on machine learning. The classification algorithm that is employed in this thesis will be described in more detail in Chapter 7.2.9. However, the type of the resolution algorithm has a bearing on the form of the feature representation. As already mentioned, we deliberately adopt a simple and well-understood algorithm, i.e. binary mention-pair classification. For this algorithm, mention pairs are represented as vectors of features which represent either properties of the potential anaphor resp. the potential antecedent itself, or properties of the *relation* between both. The two-part structure of Chapter 6.2 reflects this distinction.

6.1 Automatic Preprocessing

Each of the five dialogs in our corpus was processed by a pipeline of preprocessing components which are described in the following Chapters 6.1.2 to 6.1.6. Before that, Chapter 6.1.1 deals with the component for the detection of non-referential *it*. The de-

[34] See Chapter 7.5 for the results of experiments on manually preprocessed, *idealized* data.

scription of this preprocessing component is singled out because it is a self-contained module (described in Müller (2006)) that was completed before the rest of the work described in this thesis. For the same reason, it also features its own preprocessing (including its own sentence splitting algorithm), which is different from that used in the other preprocessing components. The entire chain of preprocessing components is depicted in Figure 10. The sentence splitting / joining processing step involves an algorithm that is much more sophisticated than the one used in the context of the detection of non-referential *it*.

6.1.1 Detection and Removal of Non-Referential *It*

On the basis of the data from the data collection that was described in Chapter 4.1, we developed a machine learning-based component to automatically detect and remove non-referential instances of *it*. These instances comprise instances of *it* that are either pleonastic, i.e. extraposed and *prop*-it, or discarded. As was described in Chapter 4.1.2, the rate of non-referential *it* in our subset of the ICSI Meeting Corpus is 37.5%. By preventing as many of these instances as possible from entering into an anaphoric relation (either as anaphor or antecedent), we expect the precision of our pronoun resolution system to improve.

6.1.1.1 Related Work

For pronoun resolution in the domain of written text, the detection of non-referential *it* has by now become a standard preprocessing step (e.g. Ng & Cardie (2002)). Paice & Husk (1987) is the first corpus-based study on the detection of non-referential *it* in written text. From examples drawn from a part of the LOB corpus (technical section), Paice & Husk (1987) create rather complex pattern-based rules which match non-referential *it* (e.g. ***SUBJECT VERB*** *it* ***STATUS*** *to* ***TASK***), and apply them to an unseen part of the corpus. They report a final success rate of 92.2% on the test corpus.

The majority of works on detecting non-referential *it* in written text uses some variant of the partly syntactic and partly lexical tests described by Lappin & Leass (1994), the first work about computational pronoun resolution to address the potential benefit of detecting non-referential *it*. Lappin & Leass (1994) mainly supply a short list of modal adjectives and cognitive verbs, as well as seven syntactic patterns indicative of non-referential *it* (e.g. *It is* ***Cogv-ed*** *that* ***S***). Like many works that treat the detection of non-referential *it* only as one of several steps of the coreference resolution process, Lappin &

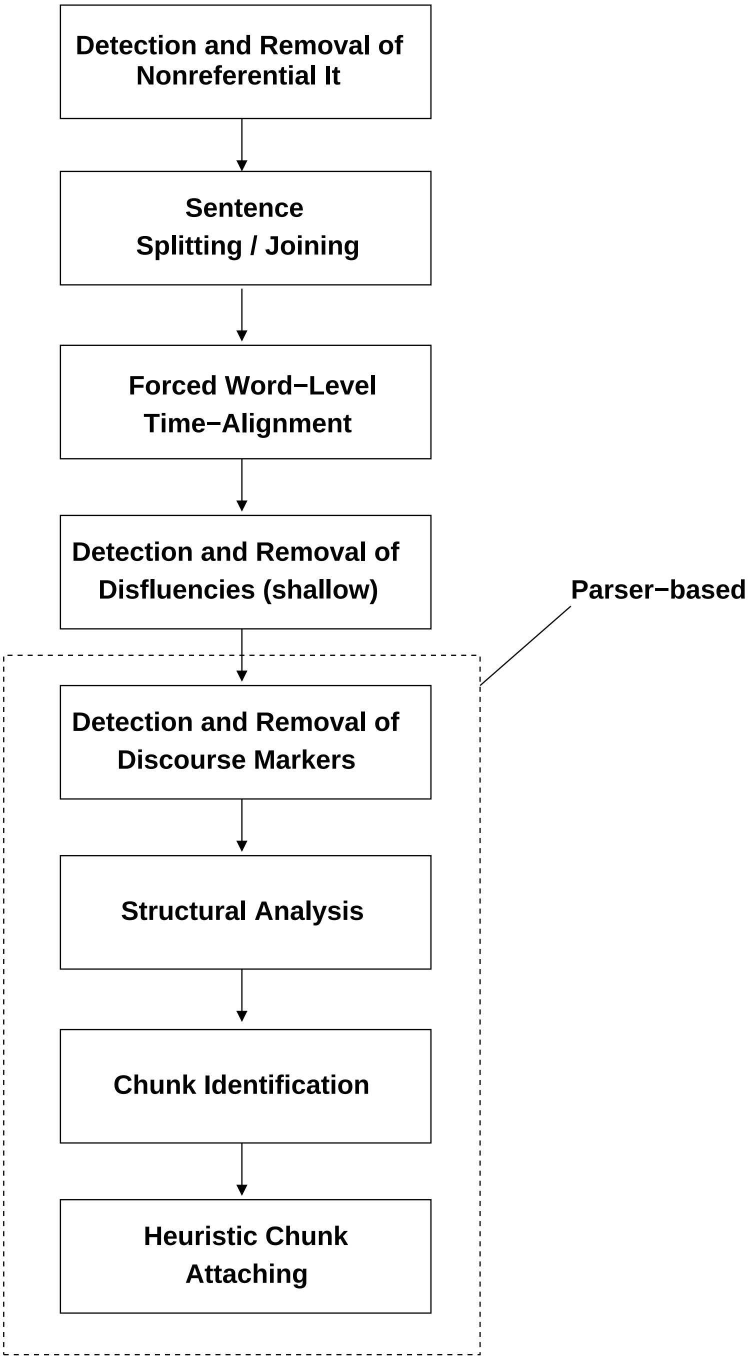

Detection and Removal of Nonreferential It
Sentence Splitting / Joining
Forced Word−Level Time−Alignment
Detection and Removal of Disfluencies (shallow)
Parser−based
Detection and Removal of Discourse Markers
Structural Analysis
Chunk Identification
Heuristic Chunk Attaching

Figure 10: Pipeline of preprocessing components.

Leass (1994) do not give any figures about the performance of this filtering method.

Dimitrov et al. (2002) modify and extend the approach of Lappin & Leass (1994) in several respects. They extend the list of modal adjectives to 86 (original: 15), and that of cognitive verbs to 22 (original: seven). They also increase the coverage of the syntactic patterns, mainly by allowing for optional adverbs at certain positions. Dimitrov et al. (2002) report performance figures for each of their syntactic patterns individually. The first thing to note is that 41.3% of the instances of non-referential *it* in their corpus do not comply with any of the patterns they use, so even if each pattern worked perfectly, the maximum recall to be reached with this method would be 58.7%. The actual recall of their system is 37.7%. Dimitrov et al. (2002) do not give any precision figures. One interesting detail is that the pattern involving the passive cognitive verb construction from Lappin & Leass (1994) accounts for only three instances in the entire corpus used by Dimitrov et al. (2002), of which only one is found.

Evans (2001) employs memory-based machine learning. He represents instances of *it* as vectors of 35 features. These features encode, among other things, information about the parts of speech and lemmata of words in the context of *it* (obtained automatically). Other features encode the presence or absence of, and the distance to, certain element sequences indicative of pleonastic *it*, such as complementizers or present participles. Some features explicitly reference structural properties of the text, like position of the *it* in its sentence, and position of the sentence in its paragraph. Sentence boundaries are also used to limit the search space for certain distance features. Evans (2001) reports a precision of 73.38% and a recall of 69.25%.

Clemente et al. (2004) work on the GENIA corpus of medical abstracts. They assume perfect preprocessing by using the manually assigned POS tags from the corpus. The features are very similar to those used by Evans (2001). Using an SVM machine learning approach, Clemente et al. (2004) obtain an accuracy of 95.5% (majority base line: approx. 56%). They do not report any precision or recall figures. Clemente et al. (2004) also perform an analysis of the relative importance of features in various settings. It turns out that features pertaining to the distance or number of complementizers following the *it* are consistently among the most important.

Finally, Boyd et al. (2005) also use a machine learning approach. They use 25 features, most of which represent syntactic patterns like *it* VERB ADJ *that*. These features are numeric, having as their value the distance from a given instance of *it* to the end of the match, if any. Pattern matching is limited to sentences, sentence breaks being identified by punctuation. Other features encode the (simplified) POS tags that surround a given instance of *it*. Like in the system of Clemente et al. (2004), all POS tag information is obtained from the corpus, so no (error-prone) automatic tagging is performed. Boyd et al. (2005) obtain a precision of 82% and a recall of 71% using a memory-based machine learning approach, and a similar precision but much lower recall (42%) using a decision tree classifier.

In summary, the best approaches for detecting non-referential *it* in written text already work reasonably well, yielding an F-measure of over 70% (Evans, 2001; Boyd et al., 2005). This can at least partly be explained by the fact that many instances are drawn from texts coming from rather stereotypical domains, like e.g. news wire text or scientific abstracts. Also, some authors make the rather unrealistic assumption of perfect POS information, and even those who do not make this assumption take advantage of the fact that automatic POS tagging is generally very good for these types of text. This is especially true in the case of complementizers (like *that*) which have been shown to be highly indicative of extraposition constructions. Structural properties of the context of *it*, including sentence boundaries and position within sentence or paragraph, are also used frequently, either as numerical features in their own right, or as means to limit the search space for pattern matching.

6.1.1.2 Features and Data Generation

We extracted all instances of *it* and the segments (i.e. speaker units) they occurred in. This produced a total of 1.017 instances, 62.5% of which were referential. Each instance was labelled as *ref* or *nonref* accordingly. Often, in particular in the presence of overlapping speech, the segments were very short, so that they did not adequately reflect the context of the *it*. Therefore, we used the segments' time information to join segments to larger units. We adopted the concept and definition of *spurt* (Shriberg et al., 2001), i.e. a sequence of speech not interrupted by any pause longer than $500ms$, and joined segments with time distances below this threshold. For each instance of *it*, features were generated mainly on the basis of this spurt. For each spurt, we performed the following preprocessing steps: First, we removed all single dashes (i.e. *interruption points*), non-lexicalised filled pauses (like *uh* and *um*), and all word fragments. This affected only the string representation of the spurt (used for pattern matching later), so the information that a certain spurt position was associated with e.g. an interruption point or a filled pause was not lost. We then ran a simple algorithm to detect direct repetitions of one to six words, where removed tokens were skipped. If a repetition was found, each token in the first occurrence was tagged as *discarded*. Finally, we also temporarily removed potential discourse markers by matching each spurt against a short list of expressions like *actually*, *you know*, *I mean*, but also *so* and *sort of*. This was done rather aggressively and without taking any context into account. The rationale for doing this was that while discourse markers do indeed convey important information to the discourse, they are not relevant for the task at hand and can thus be removed to make the (syntactic and lexical) patterns associated with non-referential *it* stand out more clearly. For each spurt thus processed, POS tags were obtained automatically with the Stanford tagger (Toutanova et al., 2003). Although this tagger is trained on written text, we used it off the shelf without any retraining.

One question we had to address was which information from the transcription we wanted to use. We assumed that information like sentence breaks or interruption points should be expected to help in the classification task at hand. On the other hand, we did not want our system to be dependent on this type of human-added information, to keep it usable for dialog corpora with a less rich manual transcription than the ICSI Meeting Corpus. Thus, we decided to create several setups which made use of this information to various degrees. Different setups differed with respect to the following two options:

-use_eos_information: This option controls the effect of explicit end-of-sentence information in the transcribed data. If this option is active, this information is used in two ways: Spurt strings are trimmed in such a way that they do not cross sentence boundaries. Also, the search space for distance features is limited to the current sentence.

-use_interruption_points: This option controls the effect of explicit interruption points. If this option is active, this information is used in a similar way as the sentence boundary information.

Each instance of *it* is represented as a sequence of features, i.e. attribute-value pairs. All of the features described in the following (cf. Table 28) were obtained fully automatically. This means that errors in the shallow feature generation methods could propagate into the model that was learned from the data. The advantage of this approach is that training and test data are homogeneous. A model trained on partly erroneous data is supposed to be more robust against similarly noisy testing data.

The first group of features consists of 21 surface syntactic patterns capturing the left and right context of *it*. Table 28 contains a selection. Each pattern is represented by a binary feature which has either the value *match* or *nomatch*. This type of pattern matching was done for two reasons: To get a simplified symbolic representation of the syntactic context of *it*, and to extract the other elements (nouns, verbs) from its predicative context. The patterns were matched using shallow (regular-expression based) methods only.

The second group of features contains lexical information about the predicative context of *it*. It includes the verb that *it* is the grammatical subject resp. object of (if any). Further features are the nouns that serve as the direct object (if *it* is subject), and the noun resp. adjective complement in cases where *it* appears in a copula construction. All these features were extracted from the patterns described above and lemmatized.

The third group of features captures the wider context of *it* through distance (in tokens) to words of certain grammatical categories, like next complementizer, next *it*, etc.

The fourth group of features contains the following: *oblique* is a binary feature encoding whether the *it* is preceeded by a preposition. *in_seemlist* is a feature that encodes whether or not the verb that *it* is the subject of appears in the list *seem, appear, look, mean, happen, sound* (from Dimitrov et al. (2002)). *Discarded* is a binary feature that encodes whether the *it* has been tagged during preprocessing as discarded due to its belonging to the *reparandum* part of a repetition.

6.1.1.3 Machine Learning

We then applied machine learning in order to build an automatic classifier for detecting non-referential instances of *it*, given a vector of features as described above. We used JRip, the WEKA (Witten & Frank, 1999) reimplementation of Ripper (Cohen, 1995), a fast and efficient rule-learner. It has been successfully applied to diverse NLP tasks, including dialog act recognition (Lendvai et al., 2003), coreference resolution (Stoyanov & Cardie, 2006), and others. A main advantage of JRip/Ripper (and of rule-learners in general) is that the rule systems that they produce are easily interpreted by humans.

All following figures were obtained by means of ten-fold cross-validation on the entire data set of 1,017 instances. Table 29 contains all results that are discussed in what follows.

In a first experiment, we did not use either of the two options described above, so that no information about interruption points or sentence boundaries was available during training or testing. With this setting, the classifier achieved a precision of 71.9%, a recall

	Syntactic Patterns	
1.	INF_it	*do it*
10.	it_BE_adj	*it was easy*
11.	it_BE_obj	*it's a simple question*
13.	it_MOD-VERBS_INF_obj	*it'll take some more time*
20.	it_VERBS_TO-INF	*it seems to be*
	Lexical Features	
22.	noun_comp	noun complement (in copula construction)
23.	adj_comp	adjective complement (in copula construction)
24.	subj_verb	verb that *it* is the subject of
25.	prep	preposition before indirect object
26.	ind_obj	indirect object of verb that *it* is subject of
27.	obj	direct object of verb that *it* is subject of
28.	obj_verb	verb that *it* is object of
	Distance Features (in tokens)	
29.	dist_to_next_adj	distance to next adjective
30.	dist_to_next_comp	distance to next complementizer (*that,if,whether*)
31.	dist_to_next_it	distance to next *it*
32.	dist_to_next_nominal	distance to next nominal
33.	dist_to_next_to	distance to next to-infinitive
34.	dist_to_previous_comp	distance to previous complementizer
35.	dist_to_previous_nominal	distance to previous nominal
	Other Features	
36.	oblique	whether *it* follows a preposition
37.	seem_list	whether subj_verb is *seem, appear, look, mean, happen, sound*
38.	discarded	whether *it* has been marked as discarded (i.e. in a repetition)

Table 28: Features for the detection of non-referential *it* (selection)

of 55.1%, and an F-measure of 62.4% for the detection of the class *non-referential*. The overall classification accuracy was 75.1%.

The advantage of using a machine learning system that produces human-readable models is that it allows direct introspection of which of the features were used, and to which effect. It turned out that the *discarded* feature was very successful. The model produced a rule that used this feature and correctly identified 83 instances of non-referential *it*, while it produced no false positives. Similarly, the *seem_list* feature alone was able to correctly identify 22 instances, producing nine false positives. The following is an example of a more complex rule involving distance features, which is also very successful (37 true positives, 16 false positives):

```
dist_to_next_to <= 8 and
```

```
dist_to_next_adj <= 4
  ==> class=nonref (53.0/16.0)
```

This rule captures the common pattern for extraposition constructions like

> **ME013**: I think it will be interesting to do other things that aren't dumb. (Bmr001)

The following rule makes use of the feature encoding the distance to the next complementizer (14 true positives, five false positives):

```
obj_verb = null and
dist_to_next_comp <= 5)
  ==> class=nonref (19.0/5.0)
```

The fact that these rules with these conditions were learned shows that the features found to be most important for the detection of non-referential *it* in written text (cf. Chapter 6.1.1.1) are also relevant for performing the same task for spoken dialog.

We then ran a second experiment in which we used sentence boundary information to restrict the scope of both the pattern matching and the distance-related features. We expected this to improve the performance of the model, as patterns should apply less generously (and thus more accurately), which could be expected to result in an increase in precision. However, the second experiment yielded a precision of only 70.1%, a recall of 57.7%, and an F-measure of 63.3% for the detection of this class. The overall accuracy was 74.9%. The system produced a mere five rules (compared to seven rules before). The model contained the identical rule using the *discarded* feature and the *seem_list* feature, with the difference that both precision and recall of the latter rule were changed: The rule now produced 23 true positives and six false positives. The slightly higher recall of the model using the sentence boundary information is mainly due to a better coverage of the rule using the features encoding the distance to the next to-infinitive and the next adjective: it now produced 57 true positives and only 30 false positives.

We then wanted to compare the contribution of the sentence breaks to that of the interruption points. We ran another experiment, using only the latter and leaving everything else unaltered. This time, the overall performance of the classifier improved considerably: precision was 80.0%, recall 60.9%, F-measure 69.2%, and overall accuracy 79.6%.

The resulting model was rather complicated, including seven complex rules. The increase in recall is mainly due to the following rule, which is not easily interpreted:[35]

```
it_S = match and
dist_to_next_nominal >= 21 and
dist_to_next_adj >= 500 and
subj_verb = null
  ==> class=nonref (116.0/31.0)
```

The considerable improvement (in particular in precision) brought about by the interruption points, and the comparatively small impact of sentence boundary information, can be explained in several ways. For instance, although sentence boundary information allows to limit both the search space for distance features and the scope of pattern matching, due to the shallow nature of preprocessing, what is *between* two sentence breaks is by no means a well-formed sentence. In that respect, it seems plausible to assume that smaller units (as delimited by interruption points) may be beneficial for precision as they give rise to fewer spurious matches. It must also be noted that interruption points do not mark *arbitrary* breaks in the flow of speech, but that they can signal important information (Heeman & Allen, 1999).

For completeness, we also ran the setting using both sentence breaks and interruption points. As can be seen in the last row of Table 29, the result was better than the one obtained when using only sentence breaks, but worse than the one using only interruption points.

	P	**R**	**F**	**% correct**
None	71.9	55.1	62.4	75.1
Sentence Breaks	70.1	57.7	63.3	74.9
Interruption Points	80.0	60.9	69.2	79.6
Both	74.2	60.4	66.6	77.3

Table 29: Classification of *It* using various information sources.

6.1.1.4 Integration

On the basis of the classifier described above, we created a filter for non-referential *it* for integration into our resolution system in the following way. For each of the five

[35] The value 500 is used as a MAX_VALUE to indicate that no match was found.

individual dialogs in our corpus, we trained a model using training data created from the remaining four dialogs. We used the setup with the best performance, i.e. the one which made use of information about interruption points. This was in line with our intention to make full use in this thesis of all information in the manual ICSI Meetincg Corpus transcript. Then, a dedicated *nonref_it* markable level was created within the annotation tool, which was populated automatically by applying the trained model to all instances of *it* in the respective test dialog, and creating a markable for each instance that was automatically classified as non-referential. The resulting markable levels (one for each dialog) could then be used in the training and testing phase of the anaphor resolution component.[36] If the filter was active, instances of *it* for which there was a markable on the markable level *nonref_it* were skipped.

6.1.2 Sentence Splitting / Joining

The next preprocessing step in our pipeline after the detection of non-referential *it* dealt with each dialog's structure of (graphemic) sentences. Each dialog was split into a sequence of graphemic sentences by using punctuation signs and capitalization available in the transcription. Due to the nature of the transcription of the ICSI Meeting Corpus, all instances of the characters **.**, **!** and **?** unambiguously indicate a sentence break. Abandoned sentences, however, lack this explicit marking. Instead, they end with a word fragment (i.e. a word ending in a dash), or an interruption point marker (i.e. a dash), like in the following example:

> **ME010**: Yeah. Yeah. No, no. There was a whole co- There was a little contract signed. It was - Yeah. (Bed017)

The sentence-final instances of word fragments and interruption point markers had to be distinguished from sentence-internal ones. This was done using the following heuristic: A word fragment or interruption point was interpreted as sentence-final if the next word after it was in uppercase, *unless* this word was identical to the last word before it, in which case it was taken to be the continuation of the utterance. This constraint was required to prevent sentence-initial repetitions (i.e. false starts like ”I - I - I think ...”) from yielding sequences of graphemic one-word sentences. Instead, by assigning them all to the same graphemic sentence, they were treated as repetitions, for which there

[36] See Chapter 7.2.8.

was a special preprocessing step (see Chapter 6.1.4 below).

The above method was applied separately for the segments of each speaker. In other words: First, all segments by speaker A in the entire dialog were processed, then all segments by speaker B, and so on. By processing the segments for each speaker separately, sentence fragments that were split by segment breaks and/or intervening segments by other speakers could be joined back together. For each identified sentence, a markable on a dedicated *sentence* markable level was created.

The number of complete resp. abandoned graphemic sentences resulting from this preprocessing step can be found in Table 30.

	Complete	**Abandoned**	**All**
Bed017	514	88 (14.62 %)	602
Bmr001	740	107 (12.63 %)	847
Bns003	627	105 (14.34 %)	732
Bro004	1028	210 (16.96 %)	1238
Bro005	1027	244 (19.20 %)	1271
Total	3936	754 (16.08 %)	4690

Table 30: Graphemic sentences.

It has to be noted that a single graphemic sentence can correspond to one or more actual sentences or clauses. In particular, a graphemic sentence that is eventually abandoned can consist of a sequence of conjoined sentences, only the last of which is abandoned, like in the following example:

> **ME010**: Because, um, you know, the standard story is that keyworks - keywords evoke frames, and the frames may well give you additional keywords or uh, if you know that - that - that a - a bunch of keywords uh, indicate a frame, then you can find documents that actually have the whole frame, rather th- than just uh, individual - (Bed017)

As the example shows, the relatively high percentage of abandoned graphemic sentences in Table 30 must not be taken to mean that a comparably high percentage of actual sentences or clauses is abandoned. Rather, the segmentation of the corpus into graphemic sentences is an artefact of the manual transcription. In the context of this thesis, the reconstruction of graphemic sentences from punctuation signs and capitalization is mainly done for obtaining reasonable units for submission to the disfluency detection (see Chapter 6.1.4) and the parser (see Chapter 6.1.5).[37]

6.1.3 Forced Time-Alignment

The ICSI Meeting Corpus does not contain word-level timing information. Timing information is only available in the form of start and end time stamps on the level of individual segments (see Chapter 3.1). In the presence of overlapping speech, the discourse order (i.e. the order in which the transcribed words appear in the transcript) is not sufficient to determine the chronological ordering of individual words. Therefore, we used a simple algorithm to map timing information from segments to the individual words. The basic idea was to distribute the known duration of the entire segment[38] approximately proportionally to all contained words. The algorithm considered only words that were actually spoken, i.e. no punctuation and no comments (i.e. markables on the *meta* markable level). Each word received a fraction of the duration of the entire segment, weighted according to the number of syllables it contained. The number of syllables was approximated as the number of graphemic vowel clusters (i.e. sequences of vowels not interrupted by a consonant) in the word. The absolute start time for each word in a segment was then calculated by adding its segment-relative start time to the absolute start time for the segment.

6.1.4 Disfluency Detection and Removal

The high percentage of speech disfluencies in the ICSI data is a major obstacle for automatically processing it. We only employed a shallow disfluency detection and removal strategy. The method was applied to the graphemic sentences previously identified.

First, non-lexicalised filled pauses (like *uh*, *um*, etc., incl. any surrounding commas), interruption points, and word fragments were removed. Like in the preprocessing for the detection of non-referential *it* (cf. Chapter 6.1.1.2), the respective tokens were only removed from the string representation of each sentence, while the information that a particular sentence position was associated with one of these disfluencies was not lost.[39] In the following example, tokens removed in this first disfluency-related preprocessing step are highlighted in grey:

[37] The only other use of this information is the feature **SameSentence**, see Chapter 6.2.2.

[38] Calculated as $segment\ end\ time - segment\ start\ time$.

[39] See e.g. the features **DistToPreviousDisfl** resp. **DistToNextDisfl** in Chapter 6.2.1.

> **ME010**: So there i- there are some - some u- uh, you know , uh, elaborations of this that you could try to put in to this structure, but I don't think it's worth it now. (Bed017)

We then ran a simple algorithm to detect direct repetitions of one to six words, which skipped those tokens that in the previous step were already marked as removed. If a repetition was found, each token in the first occurrence was removed. After the application of this algorithm, the above example was further cleansed like this:

> **ME010**: So there there are some some you know elaborations of this that you could try to put in to this structure, but I don't think it's worth it now. (Bed017)

Note how the removal of non-lexicalised filled pauses, interruption points, and word fragments in the first step makes the false starts accessible to repetition detection in the second step.

6.1.5 Parsing with Discourse Marker Detection and Removal

A syntactic analysis for every sentence was obtained by processing it with a parser trained on written text (Charniak, 2000). The choice of this particular parser was mainly motivated by its broad coverage, which was superior to that of other parsers that were also tried. One of these parsers was e.g. the Link parser, a freely available parser for Link grammar (Sleator & Temperley, 1993). This parser is particularly suited for parsing spoken, potentially disfluent language. However, its rather limited coverage turned out to be problematic for the domain of the ICSI Meeting Corpus. Therefore, a broad-coverage written text parser was chosen. Using a parser that was not trained on spoken language was not too problematic because an accurate deep structural analysis for every sentence in our corpus was not required anyway. Rather, parsing was mainly done as a means to obtain NP, VP, and adjective chunks. Output from the parser was also utilized during chunk attaching. Further details will be described in Chapter 6.1.6.

Parsing (including detection and removal of discourse markers) was done as follows: First, a cleansed string representation was created for every sentence, in which words that were marked as removed in the preceeding preprocessing steps were actually deleted. Each sentence was also matched against a list of potential discourse markers (*so*, *actually*, *like*, *you know*, *I mean*, etc.). If a sentence contained one or more matches, additional string variants were created in which the respective words (and surrounding commas, if any) were deleted. This was done exhaustively, i.e. for a sentence with two potential discourse markers, a total of four variants was created. Each of these variants (one if no potential discourse markers were identified) was then submitted to the parser. The variant with the highest probability (as determined and returned by the parser) was selected as the best one. Potential discourse markers that were deleted in that variant were marked as removed. In the example below, this method detected and removed the phrases *so*, *you know*, and *now*. Note that *you know* is clearly a discourse marker, the removal of which improves the structure of the sentence because it was embedded in the noun phrase *some elaborations*. The removal of the temporal adverb *now*, on the other hand, brings about a change in meaning which is too subtle for a coarse-grained analysis, but it doesn't improve the sentence's structure, either. The same is true for the particle *so*.

> **ME010**: So there are some you know elaborations of this that you could try to put in to this structure, but I don't think it's worth it now . (Bed017)

For comparison, the following is an example in which the above method correctly identified an instance of *you know* as *not* being a discourse marker. We first give the original, uncleansed version.

> **MN059**: And, using this and some - some uh, knowledge about the domain I think you can do some - some simple inferences. Like you know that when somebody's working about - working on - on servlets for example, he's using Java, cuz servlets are used - are written in Java. [...] (Bed017)

The final cleansed version looks like this:

> **MN059**: And, using this and some knowledge about the domain I think you can do some simple inferences. you know that when somebody's working about working on servlets for example, he's using Java, cuz servlets are used are written in Java. [...] (Bed017)

The parsing result (in the form of a Penn Treebank-like predicate-argument structure for each sentence) was stored as a set of markables on a dedicated *syntax* markable level. Each syntactic child constituent contained a markable pointer to its immediate parent constituent, and its position in its parent's left-to-right child list. This way, the entire information from the parser output could be stored.

6.1.6 Chunking and Chunk Attaching

On the basis of some of the syntactic constituents identified by the parser, markables were created on a dedicated *chunk* markable level. Chunk markables are required as potential anaphors and antecedents for pronoun resolution. The reason why they are extracted from the parser output and not from plain text is that the parser output contains additional information that is required for *attaching* different types of chunks to each other (cf. below).

6.1.6.1 NP and Adjective Chunks

NP chunk markables were created for all non-recursive NP constituents, i.e. those that did not contain other NP constituents (Abney, 1996). For Saxon genitive constructions (e.g. *Robert's idea*) and expressions with possessive pronoun determiners (*his idea*), one markable was created for the entire expression, and one for the possessive determiner part (*Robert* resp. *his*). In these cases, a special pointer attribute was used to link the full expression to the embedded determiner markable. Individual chunks for adjectives were created only for those occurring in predicative constructions. If the NP chunk was part of a prepositional phrase (PP), the string of the preposition was also stored.

6.1.6.2 VP Chunks

Based on the verbal constituents identified by the parser, VP chunk markables were also created. VP chunks were classified as one of the types **finite_verb**, **infinite_verb**, **participle**, or **modal**. For the former three types, combinations of verb and particle (e.g. *show off*) were included as phrasal verbs if a corresponding entry existed in WordNet (Fellbaum, 1998).

6.1.6.3 Chunk Attaching

Based on the tree structure that was created by the parser, NP chunks (including adjectives) and VP chunks were attached to each other using shallow heuristics.[40] The main

[40] See e.g. Buchholz (2002) for a more sophisticated approach using memory-based learning.

purpose of this was to obtain government-relations for verbs and nouns resp. grammatical role information (subject, object, etc.) for nouns, which are *not* produced by the Charniak parser. Grammatical role information is required for the calculation of some important machine learning features (see Chapter 6.2). The simple attaching algorithms used the pointer relations between syntactic parent and child constituents to traverse in a recursive top-down, left-to-right manner each constituent of type S or SBAR that was detected by the parser. NP chunks left of (finite) VP chunks were attached as probable grammatical subjects, up to three NP chunks right of VP chunks were attached as probable direct, dative or obligue objects, respectively.

The attaching was implemented by means of labelled pointer relations. E.g. the probable subject noun was attached to its verbs using a labelled relation **subject_of** (from noun to verb) resp. **subject** (from verb to noun). Complex verbal constructions like modal + infinitive were attached to each other with labelled relations like **infinitive_comp** (from modal to infinitive) resp. **infinitive_comp_of** (from infinitive to modal). The algorithms are outlined in Algorithms 2 to 5.

Algorithm 2 Algorithm `Attach_Chunks`

```
sentences ← all S-constituents returned by the parser
for all Sentences s in sentences do
  subjectNP ← rightmost immediate NP child of s
  vpChildren ← all immediate VP children of s
  for all VPs vpChild in vpChildren do
    vpHead ← head verb of vpChild
    vpHead.subject ← subjectNP
    subjectNP.subject_of ← vpHead
    ProcessVP(vpChild)
  end for
end for
```

Algorithm 3 Algorithm `ProcessVP(vp)`

```
vpHead ← head verb of vp
vpChild ← leftmost immediate VP child of vp
if vpChild exists and is not left-aligned with vp then
    AttachToVP(vpHead,vpChild)
end if
npChildren ← all immediate NP children of vp
if npChildren.length >= 1 then
    AttachToNP(vpHead,npChildren)
end if
ppChild ← leftmost immediate PP child of vp
if ppChild exists then
    npChild ← leftmost immediate NP child of ppChild
    vpHead.obl_object ← npChild
    npChild.obl_object_of ← vpHead
end if
adjChild ← leftmost immediate ADJ child of vp
if adjChild exists then
    vpHead.adj_comp ← adjChild
    adjChild.adj_comp_of ← vpHead
end if
vpChildren ← all immediate VP children of vp
for all VPs vpChild in vpChildren do
    ProcessVP(vpChild)
end for
```

Algorithm 4 Algorithm `AttachToVP(vpParentHead,vpChild)`

```
vpChildHead ← head verb of vpChild
if vpChildHead.tense = infinitive then
    vpParentHead.inf_comp ← vpChildHead
    vpChildHead.inf_comp_of ← vpParentHead
else if vpChildHead.tense = present_participle then
    vpParentHead.part1_comp ← vpChildHead
    vpChildHead.part1_comp_of ← vpParentHead
else if vpChildHead.tense = past_participle then
    vpParentHead.part2_comp ← vpChildHead
    vpChildHead.part2_comp_of ← vpParentHead
end if
```

Algorithm 5 Algorithm `AttachToNP(vpHead,npChildren)`

```
if npChildren.length = 2 then
  firstNPChild ← npChildren(0)
  secondNPChild ← npChildren(1)
  if firstNPChild = object_pronoun then
    vpHead.dative_object ← firstNPChild
    firstNPChild.dative_object_of ← vpHead
    vpHead.object ← secondNPChild
    secondNPChild.object_of ← vpHead
  else if secondNPChild = object_pronoun then
    vpHead.dative_object ← secondNPChild
    secondNPChild.dative_object_of ← vpHead
    vpHead.object ← firstNPChild
    firstNPChild.object_of ← vpHead
  else
    vpHead.dative_object ← firstNPChild
    firstNPChild.dative_object_of ← vpHead
    vpHead.object ← secondNPChild
    secondNPChild.object_of ← vpHead
  end if
else if npChildren.length = 1 then
  firstNPChild ← npChildren(0)
  if firstNPChild = object_pronoun then
    vpHead.dative_object ← firstNPChild
    firstNPChild.dative_object_of ← vpHead
  else
    vpHead.object ← firstNPChild
    firstNPChild.object_of ← vpHead
  end if
end if
for all NPs npChild in npChildren do
  vpChildren ← all immediate VP children of npChild
  for all VPs vpChild in vpChildren do
    ProcessVP(vpChild)
  end for
end for
```

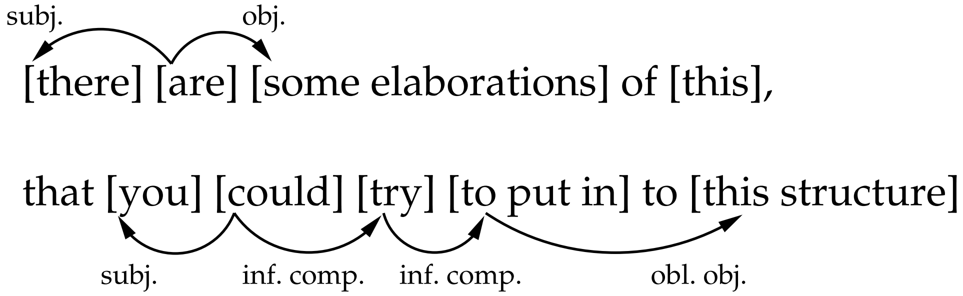

subj.
obj.
[there] [are] [some elaborations] of [this],
that [you] [could] [try] [to put in] to [this structure]
subj.
inf. comp.
inf. comp.
obl. obj.

Figure 11: Linked chunks.

Figure 11 shows a possible result of the application of Algorithms 2 to 5. In the lower part of the figure, it can be seen that some VP chunks are linked to each other to form chains of verbal expressions like modal + infinitive. In cases like these, the initial VP chunk in such a chain (*could*) is the one carrying the inflection information (the *inflected* verb). It is to this chunk that a subject, if any, is linked. The final VP chunk (the *full* verb), on the other hand, is the one carrying the lexical information proper (*to put in*), and all other arguments, if any, are linked to it. In simple, unchained VP chunks, like in the upper part of Figure 11 (*are*), the inflected and the full verb are identical. The distinction between inflected and full verb is relevant for the accurate modelling of syntactic context in terms of features (see Chapter 6.2).

6.1.6.4 Evaluation of Chunker Recall

The output of the chunker and the attributes and relations it assigns to the individual chunks are the basis for the generation of training and testing data. Markables from the *reference* markable levels receive this information by being mapped to markables on the *chunk* level. Thus, if a *reference* markable cannot be mapped one-to-one to a corresponding *chunk* markable, it will not contribute a data instance for training. More importantly, it will not be considered during testing either. This can cause the recall of the pronoun resolution to drop, because the *key* reference level will contain the respective markable nonetheless. Therefore, the recall of the chunker is an important factor for the recall of the entire system. Tables 31 to 33 contain the respective recall figures for each of the three core levels. The chunker detects approx. 90% of the reference markables. This defines an upper limit on the recall that the pronoun resolution system can reach. Soon et al. (2001) also create training and test data instances by mapping markables from the coreference annotation to markables identified by their preprocessing pipeline. They report a recall of approx. 85% for noun phrases in their MUC-6 data set.

	Annotation				Chunker abs.				Chunker rel. (%)			
	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	43	125	22	3	32	117	18	3	74.42	93.60	81.81	100.00
	193				170				88.08			
Bmr001	63	244	34	6	45	232	33	4	71.43	95.08	97.06	66.67
	347				314				90.49			
Bns003	53	173	42	1	43	160	40	1	81.13	92.49	95.24	100.00
	269				244				90.71			
Bro004	86	234	31	3	65	223	27	3	75.58	95.30	87.10	100.00
	354				318				89.83			
Bro005	83	182	21	5	67	160	20	5	80.72	87.91	95.24	100.00
	291				252				86.60			
All	328	958	150	18	252	892	138	16	76.83	93.11	92.00	88.89
	1454				1298				89.27			

Table 31: Chunker recall in core data set 2.

	Annotation				Chunker abs.				Chunker rel. (%)			
	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.	NP	PRO	VP	OTH.
Bed017	23	78	7	-	20	75	7	-	86.96	96.15	100.00	-
	108				102				94.44			
Bmr001	23	166	14	-	19	159	13	-	82.61	95.78	92.86	-
	203				191				94.09			
Bns003	25	102	18	-	21	93	18	-	84.00	91.18	100.00	-
	145				132				91.03			
Bro004	47	137	10	3	37	132	8	3	78.72	96.35	80.00	100.00
	197				180				91.37			
Bro005	45	114	10	3	37	102	9	3	82.22	89.47	90.00	100.00
	172				151				87.79			
All	163	597	59	6	134	561	55	6	82.21	93.97	93.22	100.00
	825				756				91.64			

Table 32: Chunker recall in core data set 3.

6.2 Feature Representation

In the mention-pair classification model adopted in this thesis, each data instance represents a pair of two expressions, one potential anaphor and one potential NP resp. VP antecedent. Each of these expressions is described in terms of a set of features. Features are either nominal, i.e. they take one of a set of possible values, or numeric. Some of

	Annotation				**Chunker abs.**				**Chunker rel. (%)**			
	NP	**PRO**	**VP**	**OTH.**	**NP**	**PRO**	**VP**	**OTH.**	**NP**	**PRO**	**VP**	**OTH.**
Bed017	11	40	1	-	11	37	1	-	100.00	92.50	100.00	-
		52				49				94.23		
Bmr001	11	101	2	-	10	97	1	-	90.91	96.04	50.00	-
		114				108				94.74		
Bns003	8	34	5	-	7	32	5	-	87.50	94.12	100.00	-
		47				44				93.62		
Bro004	14	44	1	-	9	41	1	-	64.29	93.18	100.00	-
		59				51				86.44		
Bro005	17	40	3	1	15	33	3	1	88.24	82.50	100.00	100.00
		61				52				85.25		
All	61	259	12	1	52	240	11	1	85.25	92.66	91.67	100.00
		333				304				91.29		

Table 33: Chunker recall in core data set 4.

the features in a data instance apply irrespective of the type of expression (anaphor, NP antecedent, VP antecedent), some apply only to nominal expressions (anaphor and NP antecedent), and some apply only to verbal expressions (VP antecedents). Table 34 provides a list of all features including the type of expression that they are applicable to. Starred features belong to the group of features referred to as *tipster* features in Chapter 7.2.5.

In addition, each data instance also encodes properties of the *relation* between both expressions. Again, some of these features apply both to pairs of anaphor and NP antecedent and to pairs of anaphor and VP antecedent, while others are restricted to one of these. Table 35 provides a list. Relations belonging to the *tipster* class are starred.

In the binary classification model, a pronoun is resolved by creating a set of candidate antecedents and searching this set for a matching candidate. This search process is mainly influenced by two factors: exclusion of candidates due to *constraints*, and selection of candidates due to *preferences* (Mitkov, 2002). Our features and relations encode information relevant to these two factors, plus more generally descriptive factors like distance etc. The column **Type** in Tables 34 and 35 classifies each feature as encoding either a *constraint* (**c**), a *preference* (**p**), or a *general* feature (**g**). Note that this can provide a rough classification only, as some features are not easily classified as either one or another class. Clear cases include the following ones: Constraints are encoded e.g. by **IncompletePredication** resp. **IncompletePredicationVP**, **Locally/GloballyRFAccessible**,

Number-, Gender-, and PersonRelation, **CoArguments**, and **ArgumentOf**. Preferences are encoded e.g. by **PredNounIdent**, **AnaPredNounMatchAnte**, **PredAdjIdent**, **PrepIdent**, **RetainForm**, **RetainGramFunc**, **LocalCenterEstablishing**, or **Full/ InflectedVerbIdent**.

As will become clear in the following, the process of feature definition for this thesis was rather unconstrained, meaning that a feature was added whenever it seemed potentially useful and reasonably operationalizable. As a result, the feature set contains a few groups of features that are potentially equivalent resp. that contain redundant features. For easy identification, these groups of features are specially marked in Tables 34 and 35.

While fully automatic pronoun resolution on spoken dialog has not yet been attempted, various machine learning approaches for other types of resolution do exist (see Chapter 5). In quite a few of these, features have been defined which can also be found in the feature set of the present thesis. For the sake of completeness, these will also be included in the following description, with references to earlier work where appropriate. However, the features that were especially devised for this thesis are more relevant here. In order for these features to be more easily identifiable, they are marked as **Novel** in Tables 34 and 35. We classify a feature as novel if it has not (to our knowledge) been used in a similar automatic pronoun resolution setting. This definition explicitly includes, but is not limited to, the numerous features that constitute operationalizations of constraints and preferences identified in the more linguistically inclined literature (see e.g. Chapters 5.1 and 5.2).

The computation of all features and relations described in the following was done fully automatically.

6.2.1 Features

Type (Ana, NPAnte, VPAnte)

The morphological type of the expression. Possible values for the anaphor are *pronoun* and *demonstrative*, possible values for NP antecedents are *noun*, *proper noun*, *pronoun*, and *demonstrative*, and possible values for VP antecedents are *infinite_verb*, *finite_verb*, *participle*, or *modal*. The value for this feature is determined on the basis of the POS tags assigned by the parser during preprocessing (Chapter 6.1.5). The motivation for this feature is to allow for a high-level separation of instances depending on the type of

Name	Ana	NP Ante	VP Ante	Type	Novel?
Type	x	x	x	g	*
Distance to prev. *uh,um*	x	x	x	g	*
Distance to prev. IP or WF					
Distance to prev. disfluency					
Distance to next *uh,um*	x	x	x	g	*
Distance to next IP or WF					
Distance to next disfluency					
Embedding in immediate clause	x	x	x	g	
Embedding in top clause					
Conj. in Passonneau-List	x	x	x	g	*
Number	x	x		g	
Gender	x	x		g	
Detailed Grammatical function	x	x		g	
Grammatical function					
Do-Object	x	x		g	
Tense of governing verb	x	x		g	*
Prep. in Paice/Husk-List	x	x		c	
Object in existential constr.	x	x		p	
Adjective-ToInfinitive CondProb*	x	x		g	*
Adjective-Complementizer CondProb*	x	x		g	*
Verb-ToInfinitive CondProb*	x	x		g	*
Verb-Complementizer CondProb*	x	x		g	*
IncompletePredication	x	x		c	*
ClauseCompType	x	x		g	*
Lemma	x			g	
Ante-Category		x	x	g	
Ante-NP size		x		g	
Ante-Person		x		g	
Ante-Determiner		x		g	
Ante-VP size			x	g	
Argument count			x	g	
Locally RF-Accessible			x	c	*
Globally RF-Accessible					
Proposition provider			x	g	*
Concept provider			x	g	*
Concept chain position			x	g	*
Existential sentence			x	p	*
Progressive CondProb*			x	g	*
Perfect CondProb*			x	g	*
IncompletePredicationVP			x	c	*

Table 34: Features for individual expressions.

Name	Ana-	Ana-	Type	Novel?
	NP Ante	VP Ante		
Word distance	x	x	g	
Temporal distance	x	x	g	*
Same sentence	x	x	g	
Same immediate clause				
Same top clause				
Same speaker	x	x	g	
Number relation	x		c	
Gender relation	x		c	
Person relation	x		c	
Predicated noun identity	x		p	*
Predicated noun matches ante	x		p	*
Predicated adjective identity	x		p	*
Preposition identity	x		p	*
Coarguments	x		c	
Retained form	x		p	
Retained grammatical function	x		p	
Establishing *local center*	x		p	*
Full verb identity	x		p	
Inflected verb identity	x		p	
NP-Adj CondProb*	x		g	
NP-Verb CondProb*	x		g	
NP distance	x		g	
Both pronouns	x		p	*
Inflected verb tense identity	x		c	*
Argument-of		x	c	
VP distance		x	g	
Tense match		x	c	*

Table 35: Features for relations between expressions.

expression involved.

DistToPreviousUh, **DistToNextUh** (Ana, NPAnte, VPAnte)

The distance in words from the expression to the closest previous resp. next word matching the regular expression ”^[Uu][hm]$”, i.e. a non-lexicalised filled pause. The default value is $10,000$. The search is limited to the graphemic sentence in which the expression appears. The distance is measured on the raw graphemic sentence, i.e. with no prior disfluency detection and removal. Just like **DistToPreviousIpWf** and **DistToNextIpWf** (cf. below), these features have not been used in computational anaphora resolution before, because previous approaches relied on data from which speech disfluencies were manually removed.

The distance to a *following* non-lexicalised filled pause can convey information about the current expression being part of an abandoned utterance, while the proximity of a *preceeding* one has been shown in the psycholinguistic literature (e.g. Arnold et al. (2003), Arnold et al. (2004)) to signal that the current expression is probably more salient.

DistToPreviousIpWf, **DistToNextIpWf** (Ana, NPAnte, VPAnte)

The distance in words from the expression to the closest previous resp. next word matching the regular expression ”(^-$|^.+-$)”. The regular expression matches single dashes (*interruption points*) and words ending in dashes (*word fragments*). The default value is $10,000$. The search is performed in and limited to the raw graphemic sentence.

DistToPreviousDisfl, **DistToNextDisfl** (Ana, NPAnte, VPAnte)

The distance in words from the expression to the closest previous resp. next disfluency, computed as

distToPreviousDisfl(EXP) = MIN(distToPreviousUh(EXP), distToPreviousIpWf(EXP))

distToNextDisfl(EXP) = MIN(distToNextUh(EXP), distToNextIpWf(EXP))

ImmediateClauseDepth, TopClauseDepth (Ana, NPAnte, VPAnte)

The depth of embedding of the expression in its immediate resp. top clause. The default value is 0. The immediate clause is the closest parent constituent of type S or SBAR detected by the parser (Chapter 6.1.5), whereas the top clause is the topmost parent constituent of this type. The depth is determined by recursively following the pointers from child to parent constituents, counting the steps until the closest resp. topmost constituent of type S or SBAR is encountered. These features are intended as a shallow approximation of the syntactic complexity of the constructions containing the anaphor resp. the antecedent. A similar feature has previously been used by Strube & Müller (2003).

ConjInPassonneauList (Ana, NPAnte, VPAnte)
Whether the immediate clause (cf. above) containing the expression is governed by one of the conjunctions appearing in the list of Passonneau (1994, p.30). Possible values are *na*, *true* and *false*. The default value is *na*. The value is *true* if the expression occurs in a clause that is governed by one of the conjunctions *after, albeit, although, as, because, before, ergo, forasmuch, how, if, inasmuch, lest, like, once, providing, since, so, then, though, till, til, until, unless, when, whence, whenever, where, whereas, whereat, whereby, wherefrom, whether, while, yet*. The value is *false* if the expression occurs in a clause that is governed by some other conjunction.
Passonneau (1994) argues that being governed by one of the listed conjunctions is a necessary (but not sufficient) requirement for a clause to be functionally independent.[41]
Thus, the feature **ConjInPassonneauList** is related to the distinction between main and subordinate clause. Passonneau (in Schiffman (1985, p.33)) previously established a connection between what she calls the *clause level* of the anaphor (and to a lesser degree also the antecedent) and its probability to be realized as *it* vs. *this* resp. *that*.

Number (Ana, NPAnte)
The number of the expression. Possible values for the anaphor are *singular* and *unknown*. Possible values for NP antecedents are *singular*, *plural* and *unknown*. The default value is *unknown*. VP antecedents have the value *na*.
Number information for nouns and proper nouns is derived directly from the syntactic category assigned by the parser. Pronouns are assigned their morphological number, except for *you*, which is ambiguous with respect to number and which is set to *unknown*. The plural demonstratives *these* and *those* are set to *plural*, whereas *this* and *that*, due to their being underspecified with respect to number (Channon, 1980), are set to the default value *unknown*.

[41] Other criteria for functional independence which are not easily operationalized include that the clause be ”semantically complete and fully specified” (Passonneau, 1994, p.21).

Gender (Ana, NPAnte)

The gender of the expression. Possible values are *masculine*, *feminine*, *human*, *neuter*, and *unknown*. The default value is *neuter*. VP antecedents have the value *na*.

More specific gender information for nouns is determined with a heuristic based on WordNet (Fellbaum, 1998). In WordNet, the synset **[*person, individual, someone, somebody, mortal, human, soul*]** is the top level parent of person-denoting synsets. For the head of each noun, we retrieve all synsets (corresponding to all possible readings). For each of these, we recursively retrieve the top level parent synset. If the parent synset of at least 50% of all readings is **[*person, individual, someone, somebody, mortal, human, soul*]**, the gender category *human* is assigned to the noun. By using the 50% threshold, the effect of rare and obscure WordNet readings is minimized.

For proper nouns, we use comprehensive lists of male and female first names and a list of surnames[42] to assign more specific gender information. If the proper noun is found in one of these lists, the respective gender category (*masculine*, *feminine*, or *human*) is assigned.

Pronouns are assigned their morphological gender, if unambiguous, or *human* if they are *I*, *me*, *you*, *we*, or *us*, else *unknown*. The default values for a pronoun can be overriden if it is the subject in a noun-copula construction where the predicated noun has a more specific gender, like in the following example:

MN015: That will be *Pause* Reuter? (Bed017)

Here, the pronoun *that* receives the gender from the proper name *Reuter*, which is automatically classified as *human* based on the list of surnames.

DetGramFunc (Ana, NPAnte)

The detailed grammatical function of the expression. Possible values are *none*, *subject*, *object*, *obl_object*, and *dat_object*. The default value is *none*. VP antecedents have the value *na*. The values are determined on the basis of the grammatical relations that were assigned during automatic chunk attaching (Chapter 6.1.6.3).

[42] Obtained from http://www.census.gov/genealogy/names/

$$gramFunc(EXP) = \begin{cases} \textit{subject} & \text{if EXP.subject\_of != <empty>} \\ \textit{object} & \text{if EXP.object\_of != <empty>} \\ \textit{obl\_object} & \text{if EXP.obl\_object\_of != <empty>} \\ \textit{dat\_object} & \text{if EXP.dat\_object\_of != <empty>} \end{cases}$$

GramFunc (Ana, NPAnte)

The simplified grammatical function of the expression. Possible values are *none*, *subject*, *object* and *other*. The default value is *none*. VP antecedents have the value *na*. The values are determined on the basis of the grammatical relations that were assigned during automatic chunking (Chapter 6.1.6).

$$gramFunc(EXP) = \begin{cases} \textit{subject} & \text{if EXP.subject\_of != <empty>} \\ \textit{object} & \text{if EXP.object\_of != <empty>} \\ \textit{other} & \text{if EXP.dat\_object\_of != <empty>} \\ \textit{other} & \text{if EXP.obl\_object\_of != <empty>} \end{cases}$$

The granularity of this feature is comparable to that used by Schiffman (1985). Byron (2003) uses an even simpler scheme which only distinguishes *subject* from *non-subject*.

DoObject (Ana, NPAnte)

Whether the expression is the object of *do*. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is *true* if the expression's **GramFunc** (cf. above) is *object* and if the verb governing it is a form of *do*. This is a simplified version of one the features used by Schiffman (1985). Byron (2003) originally also included this feature, but dropped it later for reasons of data sparsity. This feature is also used to trigger the creation of VP antecedent candidates (Chapter 7.2.3).

GovVerbTense (Ana, NPAnte)

The tense of the *inflected* (Chapter 6.1.6.3) verb governing the expression. Possible values are *undetermined*, *infinitive*, *past*, *present*, *present_participle*, *past_participle*, and *modal_conditional* (for *could*, *would*, *should*)). The default value is *undetermined*. VP antecedents have the value *na*. The values are computed on the basis of the grammatical relations that were assigned during automatic chunking (Chapter 6.1.6). This feature is intended as a very shallow way of including the temporal aspect of dialog structure into the resolution. See also the relation **TenseMatch** below.

PrepInPaiceHuskList (Ana, NPAnte)
Whether the expression is headed by one of the prepositions appearing in the list of Paice & Husk (1987, p.131). Possible values are *na*, *true* and *false*. The default value is *na*. VP antecedents also have the value *na*. The value is *true* if the expression is headed by one of the prepositions *among, before, beside, despite, from, in, near, of, onto, through, under, via, with, at, below, between, during, inside, off, outside, to, until, within, beneath, by, into, on, over, without*. The value is *false* if the expression is headed by some other preposition.
Paice & Husk (1987) use this list in their *Initial Preposition Rule* for the detection of non-referential *it*. They claim that the prepositions in their list are indicators for referential usage.[43]

ExistentialSubject (Ana, NPAnte)
Whether the expression is the object in a clause with an existential *there* as subject. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is *true* if the expression is *object*, the verb governing it is a form of *be* and the subject is existential *there*.[44]
This feature is intended to capture the same type of information as Lappin & Leass (1994)'s **Existential Emphasis** salience factor. See also the feature **ExistentialSentence** below.

AdjectiveToInfinitiveCondProb, AdjectiveComplementizerCondProb (Ana, NPAnte)
If the expression is the subject in an adjective-copula construction: The conditional probability of the predicated adjective to occur with a *to*-infinitive resp. a *that*-sentence complement. The default value is 0.000000. Both values are calculated on the basis of corpus counts in the approx. 250,000,000 word TIPSTER corpus (Harman & Liberman, 1994), using the following queries:

$$\frac{\#\,\text{it}\;\;(\text{'s}\mid\text{is}\mid\text{was}\mid\text{were})\;\;\text{ADJ}\;\;\text{to}}{\#\,\text{it}\;\;(\text{'s}\mid\text{is}\mid\text{was}\mid\text{were})\;\;\text{ADJ}}$$

[43] This feature was not identified until the filter for non-referential *it* (Chapter 6.1.1) was finished. For this reason, it has been added here.

[44] Existential *there* is detected by the parser and tagged as EX.

and

$$\frac{\#\ \text{it}\ \ (\text{'s} \mid \text{is} \mid \text{was} \mid \text{were})\ \ \text{ADJ}\ \ \text{that}}{\#\ \text{it}\ \ (\text{'s} \mid \text{is} \mid \text{was} \mid \text{were})\ \ \text{ADJ}}$$

These features are operationalizations of what Eckert & Strube (2000) call *I*- resp. *A-Incompatibility*, i.e. the semantic incompatibility of a pronoun with an individual (i.e. NP) or abstract (i.e. VP) antecedent. As Eckert & Strube (2000) note, subject pronouns in adjective-copula constructions with adjectives that are only applicable to abstract entities (like e.g. *true*, *correct*, *right*) are incompatible with concrete antecedents like *car*. According to Eckert & Strube (2000), subject pronouns in adjective-copula constructions with adjectives that are only applicable to concrete entities (like e.g. *expensive*, *tasty*) are incompatible with abstract antecedents, i.e. they cannot be discourse-deictic. These two features encode the preference of an adjective to modify an individual resp. abstract entity (in the sense of Eckert & Strube (2000)). Table 36 below gives the feature values for some selected adjectives. The column *Adjective Count* contains the raw corpus counts for the adjective alone.

Adjective	Adjective Count	ToInfinitiveCondProb	ComplementizerCondProb
true	17,975	0.004587	0.541284
correct	28,885	0.227273	0.022727
right	180,714	0.209169	0.028653
expensive	17,986	0.253165	0.006329
tasty	341	0.000000	0.000000
easy	16,671	0.603093	0.002577
great	84,604	0.004141	0.008282
edible	4,325	0.000000	0.000000
poisonous	648	0.000000	0.000000
mechanical	13,360	0.000000	0.000000

Table 36: Some examples of adjective-infinitive and -complementizer compatibility.

Note that the counts are done by mere string matching only, without taking *parts of speech* into account.

VerbToInfinitiveCondProb, VerbComplementizerCondProb (Ana, NPAnte)
If the expression is object: The conditional probability of the full verb governing the expression to occur with a *to*-infinitive resp. a *that*-sentence complement. The default value is 0.000000. Both values are calculated on the basis of corpus counts in the approx.

250,000,000 word TIPSTER corpus (Harman & Liberman, 1994), using the following queries:

$$\frac{\#\,(\text{VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING})\ \ \text{to}}{\#\,(\text{VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING})}$$

and

$$\frac{\#\,(\text{VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING})\ \ \text{that STARTER}}{\#\,(\text{VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING})}$$

Correct inflected forms are created for regular as well as irregular verbs. In the second query, STARTER stands for the regular expression ”(the | this | that | these | those | I | my | you | your | he | his | she | her | it | its | we | our | they | their)”. It is intended to match words that start a nominal expression, and is to exclude matches in which *that* is not a complementizer, but a determiner.

According to Eckert & Strube (2000), pronouns that are objects of verbs which mainly take sentence complements (like *assume*, *say*) exhibit an incompatibility with NP antecedents. Table 37 below gives the feature values for some selected verbs. The column *Verb Count* contains the raw corpus counts for any of the four verb forms (e.g. *say*, *says*, *said*, and *saying*).

Verb	Verb Count	ToInfinitiveCondProb	ComplementizerCondProb
say	2,538,034	0.003667	0.009032
assume	19,980	0.065916	0.103453
deny	34,974	0.005375	0.065620
eat	12,207	0.001884	0.000082

Table 37: Some examples of verb-infinitive and -complementizer compatibility.

Note that the counts are done by mere string matching only, without taking *parts of speech* into account.

IncompletePredication (Ana, NPAnte)

If the expression is subject and if it is governed by a form of *be*: Whether the expression is the subject in an incomplete predication. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is *true* if there is no argument (apart from the subject) attached to the verb. This feature is supposed to identify abandoned utterances that cannot be identified by simple pattern matching.

ClauseCompType (Ana, NPAnte)
The type of clause complement that the expression is an argument of. Possible values are *none*, *extrapos_subject*, *extrapos_object*, *cause*, *effect*, *manner*, *temporal*, and *local*. The default value is *none*. VP antecedents have the value *na*.
The prototypical form of relevant constructions is

It/this/that	is/was	what	(subject/object)
		because	(cause)
		why	(effect)
		how	(manner)
		when	(temporal)
		where	(local)

An example for an *extrapos_object* is

> **MN059**: And [that]'s what I then propagate back to the user, and - and try to optimize the search in this way. (Bed017)

Lemma (Ana)
The string representation of the anaphor. Possible values are *it*, *this*, and *that*. This feature allows a high-level differentiation of the anaphor type, e.g. for the implementation of different resolution strategies for pronouns vs. demonstratives.

Ante-Category (NPAnte, VPAnte)
The category of the antecedent. Possible values are *unknown*, *nominal* and *verbal*. The default value is *unknown*. NP antecedents have the value *nominal*, VP antecedents have the value *verbal*. The motivation for this feature is same as above.
The feature is simpler than the similar feature **ante_exp_type** by Strube & Müller (2003), who make an additional distinction among non-NP antecedents into VP and S antecedents. See also the **Type** feature above.

Ante-NPSize (NPAnte)
The size of the NP antecedent in terms of the number of words it contains. The default value is 1. VP antecedents have the value 0. The size of the antecedent is intended as a shallow and easily operationalized approximation of its 'importance'.

Ante-Person (NPAnte)
The grammatical person of the NP antecedent. Possible values are *first*, *second* and *third*. The default value is *third*. VP antecedents have the value *na*.

$$person(ANTE) = \begin{cases} \textit{first} & \text{if EXP} = \textit{I,me,my,we,our,us} \\ \textit{second} & \text{if EXP} = \textit{you,your} \end{cases}$$

Ante-Determiner (NPAnte)
The determiner of the NP antecedent. Possible values are *none*, *def*, *indef*, *demo*, *poss*, and *poss_np*. The default value is *none*. VP antecedents have the value *na*.

Ante-VPSize (VPAnte)
The size of the VP antecedent in terms of the number of words in the entire phrase that the VP antecedent is the head of. The default value is *1*. NP antecedents have the value *0*. The motivation for this feature is similar to that for the feature **Ante-NPSize** (cf. above).

ArgumentCount (VPAnte)
If the VP antecedent is the rightmost verbal expression, i.e. if it does not have any *verbal* complements: The number of nominal arguments attached to it, not counting the subject (if any). The default value is *0*. NP antecedents also have the value *0*. The values for this feature are determined on the basis of the grammatical relations that were assigned during automatic chunking (Chapter 6.1.6). The number of nominal arguments of a VP antecedent is used as a shallow approximation of its 'importance' (cf. **Ante-NPSize** and **Ante-VPSize**).

LocallyRFAccessible, GloballyRFAccessible (VPAnte)
Whether the phrase that the VP antecedent is the head of is right-aligned within (i.e. ends at the same position as) its closest resp. topmost parent constituent of type S or SBAR. Possible values are *true* and *false*. The default value is *false*. NP antecedents have the value *na*.
These features are inspired by Strube & Müller (2003) and try to approximate the notion of *right frontier* (Webber, 1991). Webber (1991) originally describes the accessibility of (in our terminology) VP antecedents in terms of their position in a discourse struc-

ture representation that she assumes to be tree-like. In this representation, only nodes on the rightmost branches (the right frontier of the tree) are assumed to be accessible for discourse-deictic reference. The features **LocallyRFAccessible** resp. **GloballyRFAccessible** apply the same heuristic, but in default of discourse structure analysis, they attempt it on the basis of syntactic rather than discourse structural relations. A major difference between our features and those used by Strube & Müller (2003) is that the latter employ the perfect syntactic analysis from the Penn Treebank, while our system is based on automatically obtained syntactic analyses.

PropositionProvider (VPAnte)

Whether the VP antecedent provides a proposition-like entity for discourse-deictic reference. Possible values are *true* and *false*. The default value is *false*. NP antecedents have the value *na*. The value is *true* for all VP antecedents that are of type *finite_verb* or *modal*. This feature (and the similar **ConceptProvider** below) is intended as a very rough classification of VP antecedents in terms of the abstract entities that they provide for discourse-deictic reference.

ConceptProvider (VPAnte)

Whether the VP antecedent provides a concept-like entity for discourse-deictic reference. Possible values are *true* and *false*. The default value is *false*. NP antecedents have the value *na*. The value is *true* for all VP antecedents that are not linked to any other VP chunk. For those VP antecedents that are linked to one or more other VP chunks, it is also *true* if they are of type *finite_verb* or *modal*.

ConceptChainPosition (VPAnte)

If **ConceptProvider** is *true*: The number of VP chunks between the current VP antecedent and the end of the verbal chain. The default value is *0*. NP antecedents also have the value *0*.

ExistentialSentence (VPAnte)

Whether the VP antecedent is the verb in a construction with existential *there*. Possible values are *true* and *false*. The default value is *false*. NP antecedents have the value *na*. This feature adopts the motivation of the feature **ExistentialSubject** (increase of salience) (cf. above) and applies it to VP antecedents.

ProgressiveCondProb, PerfectCondProb (VPAnte)
The conditional probability of the VP antecedent to occur in the progressive resp. perfect tense. The default value is 0.000000. Both values are calculated on the basis of corpus counts in the approx. 250,000,000 word TIPSTER corpus (Harman & Liberman, 1994), using the following queries:

$$\frac{\text{\# (I | you | he | she | it | we | they) (’m | am | ’re | are | ’s | is | was | were) VERBING}}{\text{\# VERB}}$$

and

$$\frac{\text{\# (I | you | he | she | it | we | they) (’ve | have | has) VERBED}}{\text{\# VERB}}$$

The motivation for these features is based on an observation by Eckert & Strube (2000). They observe that clauses describing states (like *Mary knows French.*) cannot serve as antecedents for discourse-deictic anaphors whose predicative context makes it clear that they require a VP referent of type **Event** (like *That happens frequently.*). Thus, knowing about the type of a VP referent might be useful for resolution. In Chapter 2.2.3, the classification of situations and different event-types by Moens & Steedman (1988) was introduced (see Table 38 below, repeated from page 34). Moens & Steedman (1988) establish a connection between the aspectual construction (mainly progressive vs. perfect) that a verb can appear in, and the type of referent it yields.

	EVENTS		STATES
	atomic	extended	
+conseq.	**CULMINATION** recognize, spot, win the race	**CULMINATED PROCESS** build a house, eat a sandwich	understand, love, know, resemble
-conseq.	**POINT** hiccup, tap, wink	**PROCESS** run, swim, walk, play the piano	

Table 38: Event-types and states, reproduced from Moens & Steedman (1988).

Moens & Steedman (1988) provide sample verbs both for **States** and for each type of **Event**. They claim that these verbs typically yield what they call *propositions* of the respective types. In the terminology used in this thesis, this can also be interpreted as the verbs sponsoring certain types of VP referents. The so-called *Aktionsart* of a verb has an influence on the grammatical *aspect* (progressive vs. perfect) of constructions in which the verb occurs. In other words: Certain aspectual constructions are only possible for verbs yielding (in our terminology) VP referents of certain types. Moens & Steedman (1988) state the following constraints:

1. A *progressive* form is only possible for verbs denoting a (culminated or nonculminated) **Process**. Violation of this constraint is claimed to be responsible for the markedness of sentences like

> # John was breaking a priceless vase at the reception.

because the progressive aspect presupposes an element of duration that is not present in the verb *break*. For the same reasons, verbs denoting **States** are generally incompatible with the *progressive*.[45]

2. A *perfect* form is only possible for verbs denoting an (atomic or extended) **Culmination**. As mentioned above, **Culminations** differ from **Points** and **Processes** in that they lead to an altered state of the world. The *perfect* form is the grammatical means to express that the consequence of this state is still relevant. Moens & Steedman (1988) provide the following example (their example 22)

> # The star has twinkled.

to illustrate the violation of this constraint: The sentence is marked because the twinkling of the star fails to bring about the relevant consequence that the *perfective* aspect presupposes. Table 39 below shows some example values. The column *Verb Count* contains the raw corpus counts for any of the three verb forms.

IncompletePredicationVP (VPAnte)
If the VP antecedent is a form of *be*: Whether the VP antecedent is an incomplete predication. Possible values are *true* and *false*. The default value is *false*. NP antecedents have

[45] But see Quirk et al. (1991, p. 202).

Verb	Verb Count	ProgressiveCondProb	PerfectCondProb
break	39, 669	0.006688	0.004376
twinkle	176	0.000000	0.000000
recognize	17, 773	0.002269	0.004538
spot	16, 414	0.000080	0.000963
win	110, 929	0.003889	0.015597
build	88, 869	0.013730	0.012884
eat	11, 397	0.019234	0.002819
hiccup	63	0.000000	0.000000
tap	5, 357	0.003877	0.004582
wink	305	0.000000	0.000000
run	86, 733	0.013794	0.003668
swim	4, 159	0.021773	0.002333
walk	21, 525	0.037142	0.002956
play	66, 122	0.016383	0.006292
understand	36, 204	0.000395	0.000461
love	19, 086	0.000280	0.001331
know	142, 843	0.000084	0.003821
resemble	2, 394	0.000000	0.000825

Table 39: Some examples of verb compatibility with progressive and perfect aspect.

the value *na*. The value is *true* if there is no argument (apart from the subject) attached to the verb. The motivation for this feature is the same as that for the feature **Incomplete Predication** (cf. above).

6.2.2 Relations

WordDistance (Ana-NPAnte, Ana-VPAnte)
The distance in words between anaphor and antecedent. The default value is $10,000$. This counts the intervening words in discourse order, i.e. in their order of occurrence in the manual segmentation. Punctuation signs and words on the *meta* level that are not actually spoken are not counted. For NP antecedents, counting starts at the last word of the expression, while for VP antecedents, counting starts at the last word of the phrase that the VP antecedent is the head of.

TempDistance (Ana-NPAnte, Ana-VPAnte)
The distance in seconds between anaphor and antecedent. The default value is $10,000$. This determines the temporal distance on the basis of the simple forced alignment described in Chapter 6.1.3. Like for the **WordDistance** feature, punctuation signs and

words on the *meta* level that are not actually spoken are not counted. For NP antecedents, the end is the last word of the expression, while for VP antecedents, the end is the last word of the phrase that the VP antecedent is the head of.

SameSentence (Ana-NPAnte, Ana-VPAnte)

Whether anaphor and antecedent occur in the same graphemic sentence. Possible values are *true* and *false*. The default value is *false*. This feature is intended to capture the proximity of two expressions in a way that is independent of the word or temporal distance (cf. above).

SameImmediateClause, SameTopClause (Ana-NPAnte, Ana-VPAnte)

Whether anaphor and antecedent occur in the same immediate resp. top clause. Possible values are *true* and *false*. The default value is *false*. Clause structure is obtained automatically on the basis of the parser output. The motivation for this feature is the same as that for the feature **SameSentence** (cf. above).

SameSpeaker (Ana-NPAnte, Ana-VPAnte)

Whether anaphor and antecedent belong to (not necessarily adjacent) utterances from the same speaker. Possible values are *true* and *false*. The default value is *false*.

NumberRelation (Ana-NPAnte)

The relation between the anaphor's and NP antecedent's **Number** features. Possible values are *ident*, *compatible*, and *incompatible*. The default value is *incompatible*. VP antecedents have the value *na*.

$$numRel(ANA, ANTE) = \begin{cases} \textit{ident} & \text{if num(ANA) = num(ANTE)} \\ \textit{comp} & \text{if num(ANA) = unknown or num(ANTE) = unknown} \\ \textit{incomp} & \text{if num(ANA) != num(ANTE)} \end{cases}$$

GenderRelation (Ana-NPAnte)

The relation between the anaphor's and NP antecedent's **Gender** features. Possible values are *ident*, *compatible*, and *incompatible*. The default value is *incompatible*. VP antecedents have the value *na*.

$$genRel(ANA, ANTE) = \begin{cases} \textit{ident} & \text{if gender(ANA) = gender(ANTE)} \\ \textit{comp} & \text{if gender(ANA) = unknown or gender(ANTE) = unknown} \\ \textit{comp} & \text{if gender(ANA) = human and gender(ANTE) = masculine} \\ \textit{comp} & \text{if gender(ANA) = masculine and gender(ANTE) = human} \\ \textit{comp} & \text{if gender(ANA) = human and gender(ANTE) = feminine} \\ \textit{comp} & \text{if gender(ANA) = feminine and gender(ANTE) = human} \\ \textit{incomp} & \text{if gender(ANA) != gender(ANTE)} \end{cases}$$

PersonRelation (Ana-NPAnte)
The relation between the anaphor's and NP antecedent's **Person** feature. Possible values are *ident* and *incompatible*. The default value is *incompatible*. VP antecedents have the value *na*.

$$personRel(ANA, ANTE) = \begin{cases} \textit{ident} & \text{if person(ANA) = person(ANTE)} \\ \textit{incomp} & \text{if person(ANA) != person(ANTE)} \end{cases}$$

PredNounIdent (Ana-NPAnte)
If both anaphor and NP antecedent are subject in noun-copula constructions: Whether the predicated noun is identical. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.
For this feature (as well as for **AnaPredNounMatchAnte**, cf. below), the comparison is done using only the lemmatized head (in case of compounds or multi-word expressions) of the resp. expressions. Lemmatization is done with the lemmatizing function of the TreeTagger (Schmid, 1994).

AnaPredNounMatchAnte (Ana-NPAnte)
If the anaphor is subject in a noun-copula constructions: Whether the predicated noun matches the antecedent. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

PredAdjIdent (Ana-NPAnte)

If both anaphor and NP antecedent are subject in adjective-copula constructions: Whether the predicated adjective is identical. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

PrepIdent (Ana-NPAnte)

Whether both anaphor and NP antecedent appear in prepositional phrases beginning with the same preposition. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

CoArguments (Ana-NPAnte)

Whether anaphor and NP antecedent are co-arguments, i.e. arguments (subject, object, or other) of the same verb token. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is determined on the basis of the grammatical relations that were assigned during automatic chunking (Chapter 6.1.6).

This feature encodes an approximation of Principle B of Binding Theory (Chomsky, 1981), which states that a non-reflexive pronoun must not be bound to an NP which c-commands it. In the context of this thesis, we are only dealing with the non-reflexives *it*, *this*, and *that*. Therefore, a relation of co-argumenthood is an unambiguous sign of anaphor and antecedent not being coreferent.

RetainForm (Ana-NPAnte)

Whether anaphor and NP antecedent consist of the same string (i.e. *it*, *this*, or *that*). Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

RetainGramFunc (Ana-NPAnte)

Whether anaphor and NP antecedent have the same grammatical function. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is *true* if both anaphor and NP antecedent have the same grammatical function of either *subject*, *object*, or *other*. If the grammatical functions are different or if both expressions have the grammatical function *none*, the value is *false*. This feature encodes what is sometimes referred to as *parallel function preference* (Kertz et al., 2006). It states that a pronoun with a particular grammatical function preferrably has an antecedent with the same function.

LocalCenterEstablishing (Ana-NPAnte)

Whether both anaphor and NP antecedent are adjacent instances of *it* that are uttered by the same speaker and that are both either subject or non-subject. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

The notion of *local center* is based on Passonneau (1991). The notion of *s-adjacency* is originally defined with respect to the discourse segment in which both pronouns appear. In default of discourse segmentation in our corpus, we simply approximate it with *actual* adjacency (i.e. no intervening instances of *it*).

FullVerbIdent (Ana-NPAnte)

Whether the lemmata of the full verbs governing anaphor and NP antecedent are identical. Possible values are *true*, *true_be* and *false*. The default value is *false*. VP antecedents have the value *na*. As described in Chapter 6.1.6.3, the full verb is the rightmost verb in a chain of verbal expressions, i.e. the one carrying the lexical information. The value is *true_be* if both verbs are forms of *be*, else *true* if both verbs are identical and non-empty.

This feature encodes information that is similar to the *association history* of Rocha (1999) (Chapter 5.2.2). We distinguish between *true* and *true_be* in order to counter the prevalence of *be* in favour of more rare (and thus more discriminative) verbs.

InflectedVerbIdent (Ana-NPAnte)

Whether the lemmata of the inflected verbs governing anaphor and NP antecedent are identical. Possible values are *true*, *true_be* and *false*. The default value is *false*. VP antecedents have the value *na*. As described in Chapter 6.1.6.3, the inflected verb is the leftmost verb in a chain of verbal expressions. In the most simple case when there is only one verbal chunk, the full verb and the inflected verb are identical. The value of this feature is *true_be* if both verbs are forms of *be*, else *true* if both verbs are identical and non-empty.

AnaPredAdjCondProb (Ana-NPAnte)

If the anaphor is the subject in an adjective-copula construction: The conditional probability of the predicated adjective to appear as a modifier or predicate of the NP antecedent (Lapata et al., 1999). The default value is 0.000000. The value is calculated on the basis of corpus counts in the approx. 250,000,000 word TIPSTER corpus (Harman & Liberman, 1994), using the following query:

$$\frac{\begin{array}{c}\#\ \text{ADJ (ANTE} \mid \text{ANTES)} + \\ \#\ \text{ANTE (is} \mid \text{was) ADJ} + \\ \#\ \text{ANTES (are} \mid \text{were) ADJ}\end{array}}{\#\ \text{ADJ}}$$

Table 40 below contains the **AnaPredAdjCondProb** values for some pairs of adjectives and nouns (taken from Lapata et al. (1999)). Each adjective is paired with three nouns, and the order of the nouns in the table follows the order by Lapata et al. (1999). In their list, the first noun is the one with the highest co-occurrence frequency, the second the one with medium and the third the one with low co-occurrence frequency. It can be seen in Table 40 that the **AnaPredAdjCondProb** feature replicates this ordering to some extent by always returning the highest value for the first noun in each triple, while in three out of four cases it is less discriminative between the second and third one.

Adjective	Adjective Count	Noun	Noun Count	AnaPredAdjCondProb
hungry	$2,149$	**animal**	$19,127$	0.000931
		pleasure	$2,025$	0.000000
		application	$44,638$	0.000000
guilty	$15,411$	**verdict**	$6,226$	0.017520
		secret	$16,059$	0.000000
		cat	$3,756$	0.000000
temporary	$10,586$	**job**	$77,504$	0.007652
		post	$34,503$	0.000472
		cap	$9,097$	0.000000
naughty	125	**girl**	$21,221$	0.024000
		dog	$12,095$	0.000000
		lunch	$5,751$	0.000000

Table 40: Some examples of adjective-noun compatibility.

AnaGovVerbAnteCondProb (Ana-NPAnte)
If the anaphor is subject or object: The conditional probability of the NP antecedent to appear as an argument of the verb governing the anaphor. The default value is 0.00000. The value is calculated on the basis of corpus counts in the approx. 250,000,000 word TIPSTER corpus (Harman & Liberman, 1994). If the anaphor is subject, the following query is used:

$$\frac{\begin{array}{c}\#\ \text{ANTE (VERBS} \mid \text{VERBED)}+\\ \#\ \text{ANTE (is} \mid \text{was) VERBING}+\\ \#\ \text{ANTES (VERB} \mid \text{VERBED)}+\\ \#\ \text{ANTES (are} \mid \text{were) VERBING}\end{array}}{\#\ \text{(ANTE} \mid \text{ANTES)}}$$

If the anaphor is object, the query looks like this:

$$\frac{\begin{array}{c}\#\ \text{(VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING) (}\varnothing \mid \text{a} \mid \text{an} \mid \text{the} \mid \text{this} \mid \text{that) ANTE}+\\ \#\ \text{(VERB} \mid \text{VERBS} \mid \text{VERBED} \mid \text{VERBING) (}\varnothing \mid \text{the} \mid \text{these} \mid \text{those) ANTES}\end{array}}{\#\ \text{(ANTE} \mid \text{ANTES)}}$$

This feature is similar to that used by Yang et al. (2005). Statistical features for quantifying the compatibility of nouns and verbs by means of predicate-argument counts have also been used by Dagan et al. (1995), Kehler et al. (2004), and Bean & Riloff (2004). The utility of this type of knowledge for pronoun resolution is not undisputed. Kehler et al. (2004) find that it fails to bring about a significant improvement in their pronoun resolution system. They come to the conclusion that predicate-argument statistics are either unnecessary, because most pronouns can be resolved on the basis of morphosyntactic cues alone, or insufficient, because difficult pronouns require world knowledge that mere corpus counts fail to represent. Bean & Riloff (2004), on the other hand, state that automatically acquired statistics about predicate-argument combinations are an important knowledge source for noun phrase and in particular pronominal coreference resolution. The latter observation they attribute to the fact that anaphoric noun phrases in themselves contain sufficient lexical information for their resolution, while this information is not available in pronouns in isolation. Therefore, pronouns are more likely to benefit from information supplied by their (predicative) context.

NPDistance (Ana-NPAnte)

The number of intervening NPs between anaphor and NP antecedent, counting all NPs. The default value is $10,000$. Counting is done in global processing order, i.e. discourse order.

BothPronouns (Ana-NPAnte)
Whether both anaphor and NP antecedent are (not necessarily the same) pronouns (i.e. *it*, *this*, or *that*). Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*.

IdentTense (Ana-NPAnte)
Whether the inflected verbs governing anaphor and NP antecedent have the same tense. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. In analogy to the feature **GovVerbTense** (cf. above), this feature tries to capture in a very shallow manner what could be called the temporal compatibility of the verbal constructions in which anaphor and NP antecedent appear.

ArgumentOf (Ana-VPAnte)
Whether the anaphor is an argument (i.e. subject, object, or other) of the VP antecedent. Possible values are *true* and *false*. The default value is *false*. VP antecedents have the value *na*. The value is determined on the basis of the grammatical relations that were assigned during automatic chunking (cf. Chapter 6.1.6).
This feature is similar to the feature **CoArguments** (cf. above) in that it is an unambiguous indicator of anaphor and VP antecedent not being coreferent.

VPDistance (Ana-VPAnte)
The number of intervening VPs between anaphor and VP antecedent, counting all VPs. The default value is $10,000$. Counting is done in global processing order, i.e. discourse order.

TenseMatch (Ana-VPAnte)
Whether the tense of the inflected verb governing the anaphor is identical to the tense of the VP antecedent. Possible values are *true* and *false*. The default value is *false*. NP antecedents have the value *na*. The motivation for this feature is the same as that for **IdentTense** (cf. above).

6.3 Chapter Summary

This chapter consisted of two major parts, each of which described a different aspect of the implementation of our spoken dialog pronoun resolution system. The first part was more technical in nature. It dealt with how the annotated corpus was preprocessed in order to extract structural and semantic information. This information is required for the modelling of the pronoun resolution task, described in the second part (cf. below). Despite the fact that the individual preprocessing steps are as diverse as POS-tagging, sentence segmentation, time-aligning, disfluency detection, parsing, chunking, and chunk attaching, they all share one characteristic: They are performed in a rather shallow, heuristic-based manner. This is because there currently are no available components for performing the diverse tasks with a reasonable robustness and coverage *on spoken language*. Exceptions to this are the POS-tagger (Toutanova et al., 2003) (only used for preprocessing in Chapter 6.1.1.2) and in particular the parser (Charniak, 2000), which are not heuristics-based, but full-blown applications. However, even these could not be used to full capacity, because they were trained on and optimized for the processing of written text. In default of available resources, improving the performance of these applications by retraining them on spoken language data was not feasible either. As a result, the information obtained by means of preprocessing – and consequentially the features and relations that were computed on the basis of this information – contain an unknown amount of noise.

The second part of this chapter described in some detail the set of descriptive features and relations that are used to model the pronoun resolution process. The set contains some fairly standard features describing e.g. distances between anaphors and antecedents, and their (in)compatibility in terms of number, gender, and person agreement. There are also quite a few novel features which have not been used for computational pronoun resolution before. Some of these novel features encode different forms of parallelism between anaphors and antecedents. Another interesting group of features are the *tipster* features, a number of numerical features representing conditional probabilities for several phenomena on the basis of counts on the Tipster corpus. Among the *tipster* features are some which quantify the semantic compatibility of verb-noun and adjective-noun pairs in the common way by means of predicate-argument and noun-modifier statistics. Other *tipster* features, like the ones based on insights of Moens & Steedman (1988) and Eckert & Strube (2000), are new and have not been used before. They are mainly intended to help in the resolution of discourse-deictic pronouns. Some of the features described in this chapter are potentially redundant, i.e. they perform very similar tasks resp. encode very similar information. While these features have been identified, no attempt has been made to reduce the feature set by removing them. In the next chapter, the features will be put to use for machine-learning based classification.

7 Experiments and Results

This chapter contains detailed descriptions of a series of machine learning experiments that we performed in order to build a system for automatically resolving instances of *it*, *this* and *that*. As already mentioned, this thesis differs from many other works in that it employs anaphora resolution as a means to perform a function within a practically usable system. In this context, the commonly applied coreference evaluation measures are inapplicable. Therefore, in Chapter 7.1 we begin by defining the functional evaluation measure that is to be used in this thesis. Note, however, that this evaluation measure will only be applied in an intrinsic fashion: An extrinsic evaluation of our system, i.e. the determination of its contribution to an extractive summarization system, is beyond the scope of this thesis (cf. Chapter 8).

As described in Chapter 5, fully automatic anaphora resolution for spoken multi-party dialog has not yet been attempted. For two-party dialog, the situation is only slightly different, with two implemented but not fully automatic systems existing (Strube & Müller, 2003; Byron, 2004). As a result, some of the parameters of the training data generation and testing phases of spoken dialog anaphora resolution are more or less unexplored. Among these parameters are e.g. the order in which the expressions in the dialog are processed (discourse vs. chronological order), the antecedent search depth, i.e. the maximal distance between antecedent and anaphor, and the degree of instance pre-filtering, i.e. which antecedent-anaphor pairs are systematically excluded due to assumed incompatibility in number or gender. In Chapter 7.2, all these parameters will be described in some detail. In our resolution experiments, we will systematically vary them to determine the setting yielding the best overall result.

Chapter 7.3 contains details about how the training data generation was performed, while Chapter 7.4 describes the actual machine learning experiments for which this data was used. Our experiments will be targeted at the task of finding NP and VP antecedents for instances of *it*, *this*, and *that*. All experiments will be performed in a realistic setting, i.e. on the basis of fully automatic preprocessing. In contrast, Chapter 7.5 describes experiments in which some preprocessing (the detection of non-referential *it*, disfluency detection and removal) is performed manually. The aim of these experiments with idealized data is to get a better idea of the performance of the actual anaphora resolution by eliminating at least some of the error sources in the preprocessing.

7.1 Evaluation Measures: MUC-Style vs. 'Functional' Evaluation

Evaluation of anaphora resolution is usually performed by comparing the resolution system output (often referred to as the *response*) to a solution that is known to be correct (the *key*), and by quantifying the degree of (dis-)similarity between both.

For pronoun resolution, the situation is normally complicated by the fact that *key* and *response* consist of sets of expressions. The *key* is the set of sets of all mentions that actually *are* coreferent, and the *response* is the set of sets of all mentions that some system *classified* as coreferent. Thus, except in simple cases where there is only exactly one antecedent for a given anaphor, it is difficult to state for a given anaphor whether it has been processed correctly or incorrectly, simply because there can be several correct antecedents. Based on this insight, Vilain et al. (1995) developed a coreference evaluation scheme for use within the MUC-6 project. The scheme quantifies the (dis-)similarity between two coreference annotations, i.e. two sets of coreference sets, by determining the number of coreference links that minimally have to be added or removed in order to make both annotations identical. For a considerable time, this scheme was the standard coreference evaluation measure, although it was also much critizised (Bagga & Baldwin, 1998; Popescu-Belis & Robba, 1998; Trouilleux et al., 2000). The scheme by Vilain et al. (1995) produces good scores for coreference resolution systems that maximize correct and minimize incorrect links. In doing so, it does not take into account the nature of the linked expressions.

Other than in the majority of works in the literature, anaphora resolution in the context of this thesis is not done as an end in itself, but as a preprocessing step that is to serve a function in an actual application. As was described in the introduction (Chapter 1), one of the possible functions is the expansion (through substitution) of pronominal expressions into lexically explicit nominal or verbal expressions. As Bagga & Baldwin (1998) and Stuckardt (2003) point out, in an application-oriented setting like this, not all anaphoric links are equally important: If a pronoun is resolved to an anaphoric chain that contains only pronouns, this resolution can be treated as neutral because it has no application-level effect due to the lack of a non-pronominal expression that the pronoun can be substituted with. This is true independently of the question whether or not the respective pronoun was resolved correctly, i.e. whether or not one or more of the pronouns in the anaphoric chain actually are the correct antecedents. Figure 12 shows an example. The bold lines represent the correct anaphoric links in the *key*, while the dashed lines represent the links in the *response*. The algorithm of Vilain et al. (1995)

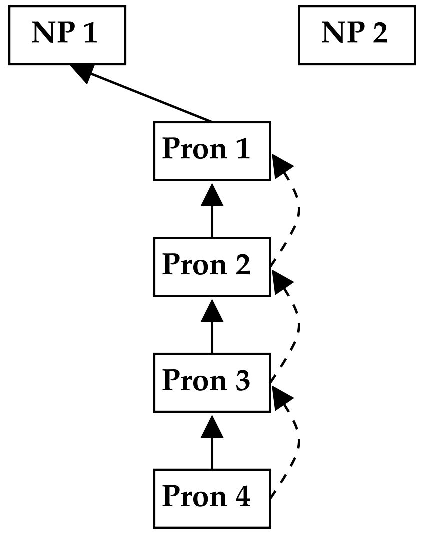


Figure 12: Unresolved anaphoric pronoun chain.

yields a recall of 75% because three out of four anaphoric links have been detected. On a functional level, however, none of the pronouns has really been resolved.

Likewise, if a pronoun is resolved to an anaphoric chain with the wrong non-pronominal chain-initial expression, this resolution has to be treated as wrong. This is true regardless of the question whether or not one or more of the other pronouns actually are the correct antecedents. Figure 13 shows an example. Here, the algorithm of Vilain et al. (1995) yields a precision of 75% because three out of four anaphoric links are correct. On a functional level, however, all four pronouns are resolved incorrectly, because the wrong chain-initial antecedent is propagated down the entire chain, leading to four incorrect substitutions.

For the present task, therefore, the evaluation scheme of Vilain et al. (1995) is inappropriate because it fails to make the distinction between pronominal and non-pronominal antecedents. Instead, we calculate precision and recall as follows:

$$Precision = \frac{Correctly\ resolved\ pronouns}{All\ resolved\ pronouns}$$

and

$$Recall = \frac{Correctly\ resolved\ pronouns}{All\ resolvable\ pronouns}$$

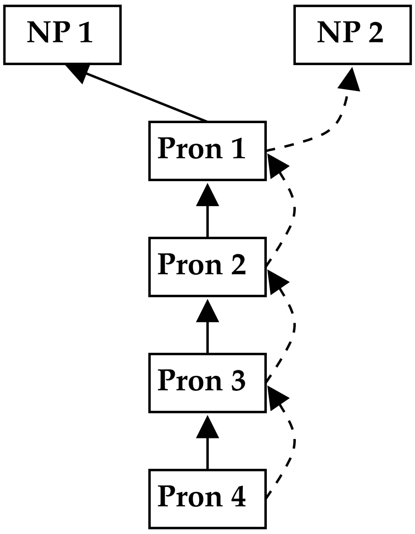


Figure 13: Incorrectly Resolved anaphoric pronoun chain.

where

Correctly resolved pronouns = number of pronouns in the *response* that are linked (directly or transitively) to the correct non-pronominal antecedent,

All resolved pronouns = number of pronouns in the *response* that are linked (directly or transitively) to a non-pronominal antecedent, and

All resolvable pronouns = number of pronouns in the *key* that are linked (directly or transitively) to a non-pronominal antecedent.

Precision, recall and F-measure are reported for NP and VP antecedents individually, and for both types of antecedents taken together (ALL). The different figures are obtained by only considering anaphoric chains with initial antecedents of the respective type when computing precision, recall, and F-measure.

7.2 Experimental Parameters

The process of training data generation, training, and testing is mainly controlled by a couple of parameters. Some of these are particular to the task of spoken language anaphora resolution, while others are also common for the task in written language.

7.2.1 Data Set

In contrast to other works on anaphora or coreference resolution which use a manually created gold standard data set (like the MUC or ACE data sets), the data basis for this thesis consists of three distinct core data sets. As was described in Chapter 4.2.4, these were created on the basis of majority decisions from the annotations of four individual annotators. Based on the assumption that an anaphoric link is the more plausible the more annotators identify it, the three sets differ with respect not only to size, but also with respect to the reliability of the links contained in them. Table 15 on page 86 (repeated as Table 41 below) contains for each dialog the number of links in the respective core data set.

	core data set 2	core data set 3	core data set 4
Bed017	116	62	28
Bmr001	229	132	69
Bns003	164	82	27
Bro004	212	111	32
Bro005	170	95	32
Σ	891	482	188

Table 41: No. of immediate links in three core data sets.

For training, a larger data set is normally to be preferred because it provides more data. On the other hand, the plausibility of the links in e.g. the core data set 2 is not optimal since the number of annotators that agree on them (i.e. two) is the same as the number of annotators that do *not* agree on them. In this respect, a data set like core data set 4 is to be preferred since it only contains links that all annotators agree upon. This, however, comes at the price of a very small data set, since core data set 4 is less than 25% the size of core data set 2. In our experiments, we trained different models using all three core data sets. This way the effect of training data quality and quantity on resolution can be studied. For testing, in contrast, one and the same data set was used throughout in order to make the different results comparable. We always used core data set 3 as evaluation key, because we think it offers a good tradeoff between sufficient size on the one and and plausibility of the contained links on the other hand. Also, the antecedent-anaphor pairs in this data set are theoretically sound since the number of annotators that agree on them is larger than the number of annotators that do not agree on them.

Related to the choice of a training and test data set is the question of the appropriate antecedent search depth, i.e. the maximum distance between an anaphor and its NP or VP antecedent. This parameter is of crucial importance for both training data generation and testing. Due to the combinatory nature of training data generation in the binary classification paradigm (cf. below), if the maximum distance is chosen too large, too many irrelevant antecedent-anaphor pairs will be created, making worse the problem of data set skewness. If it is chosen too small, a considerable number of relevant pairs might be lost. Therefore, a search depth limit for training data generation needs to be defined. If the distance between antecedent and anaphor in the current instance exceeds this limit, training instance generation will terminate for the current anaphor.[46]

For testing, limiting the maximum search depth is equally important: If too many antecedent candidates are considered for a given anaphor, the chances for errors increase. At the same time, if too few are considered, the anaphor may remain unresolved or resolved to the wrong antecedent.

Just like for the calculations of average distances (Chapter 4.2.5.3), we used the timestamps assigned in the course of the simple forced alignment described in Chapter 6.1.3 to determine the distance between anaphor and antecedent. We used the following different maximal NP and VP antecedent search depths for the three different core data sets.

Core Data Set	**NP Antecedent**	**VP Antecedent**
2	13 sec.	7 sec.
3	9 sec.	7 sec.
4	6 sec.	3 sec.

Table 42: Maximal antecedent search depths.

The distances are estimated as *average distance plus one standard deviation*, on the basis of the average distances determined in Chapter 4.2.5.3.

In anaphor resolution in written text, the antecedent search depth is often defined in terms of graphemic sentences (often the last two sentences, e.g. Yang et al. (2003)). In spoken language, this criterion is not available. The graphemic sentences that were automatically detected during preprocessing (Chapter 6.1.2) on the basis of the manual

[46] This limitation will only be effective for non-referential anaphors, i.e. those that do not give rise to a positive instance, and for referential ones whose antecedent actually exceeds the distance limit. For all other cases, training instances will only be generated up to and including the most recent antecedent (cf. Chapter 7.3).

transcription do not appear suitable because they often represent rather arbitrary decisions. In their work on spoken two-party dialog, Strube & Müller (2003) use the last two complete, non-backchannel utterances to limit the generation of pairs with (in our terminology) VP antecedents. They derive the necessary information about the nature of an utterance (i.e. whether it is abandoned) from the Penn Treebank annotation. In our corpus, similarly reliable information is not available.

We always use the same distances for training and for testing (except for the baseline, cf. below). Since all testing is done on core data set 3, this means that the maximal search depth that is used for testing matches the one that was used for training only in one out of three cases. However, we argue that using the same distances during training and testing is essential because the classifier is optimized for the distance range that it encountered during training. Confronting it with significantly different instances during testing will probably lead to poor performance.

In the baseline system, no training is involved. Therefore, instead of using three different classifiers trained on three different data sets, we run the baseline classifier with three different maximal antecedent search depths.

7.2.2 Oversampling

It has been noted as early as in McCarthy & Lehnert (1995) that binary-classification anaphora resolution is a machine learning task that is troubled with highly unbalanced (or skewed) data sets. For binary classification, the class distribution of a data set is skewed if one class is represented only by very few instances and the other by the majority of instances. For binary-classification anaphora resolution, the rare class is the one comprising pairs of anaphors and their actual antecedents, while the majority class consists of cases where antecedent and anaphor do not stand in this relation. Often, the underrepresented class is the more interesting one, i.e. the one that is to be detected. This is also the case in anaphora resolution. In practical machine learning, the problem is that learners often fail to build models that properly represent the minority class, with the result that the recall for this class is very low. A common reason for this is that many learners optimize accuracy, i.e. the proportion of correct classifications. If the majority class constitutes e.g. 90% of all instances, a learner will already achieve an accuracy of 90% by simply returning the majority class for every instance in the test data set, at the price of not finding even a single instance of the more interesting minority class.

A couple of measures have been proposed to counter the prevalence of the majority class and to increase the recall for the more interesting class (see e.g. Japkowicz & Stephen (2002) for an overview). The most common measures are negative under- resp. positive oversampling. As the name suggests, undersampling attempts to make the class distribution more balanced by removing instances of the majority class, while oversampling does so by adding instances of the minority class. In our experiments, we performed three-fold and six-fold positive oversampling. Active selection of positive instances for duplication was not performed. Instead, all positive instances were duplicated according to the respective oversampling rate.

It is common in machine learning to apply oversampling (or resampling in general) until all classes are balanced (Japkowicz & Stephen, 2002). However, this appears inappropriate for machine-learning based anaphora resolution because it might induce into the learner a bias towards the positive class which is as unwarranted as a bias towards the negative class.[47] Therefore, we used two different oversampling rates (see Chapter 7.3 for details). An oversampling rate of three produces a training data set in which the positive class is still underrepresented, while an oversampling rate of six produces a training data set in which both classes are approximately balanced. No oversampling was performed on the test data.

7.2.3 Antecedent Types: NP and VP

For the resolution of discourse-deictic anaphors, VP antecedent candidates have to be considered during data generation and testing. In their approach, Strube & Müller (2003) consider VP antecedents for all instances of *it* and *that*. Byron (2004) implements the difference between individual and potentially discourse-deictic anaphors by using different antecedent search strategies. Her algorithm preferrably treats demonstratives as discourse-deictic and pronouns as individual. It is a common observation that demonstratives (in particular *that*) are preferred over *it* for discourse-deictic reference (Schiffman, 1985; Webber, 1991; Asher, 1993; Eckert & Strube, 2000; Byron, 2004; Poesio & Artstein, 2005b).[48] This preference can also be observed in our data (cf. Tables 18 (page 92), 20 (page 93), and 22 (page 94): In the core data sets 2, 3, and 4, the rate of VP antecedents that are anaphorically referred to by *that* is 61.33%, 69.49% and 75.00%, respectively.

In line with this observation, we also create VP antecedent candidates for *that*-anaphors

[47] But see Hoste (2005).

[48] See Chapter 2.3 for a linguistically motivated explanation of this fact.

only. This decision is both empirically and linguistically justified. In addition, we also create VP antecedent candidates for all anaphors that appear in the direct object position of a form of the verb *do*. It is generally agreed (e.g. Schiffman (1985), Eckert & Strube (2000)) that this is a strong indication for the anaphor to be discourse-deictic.[49]

7.2.4 Instance Filtering

Incompatibility of anaphor and antecedent in the categories number, gender and person is a commonly applied filter for the creation of nominal antecedent-anaphor pairs in automatic resolution (Mitkov, 2002). By excluding incompatible pairs during training, the number of negative instances is kept low, which is desirable in the binary classification paradigm since it counters class skewness. Consequently, incompatible pairs are also commonly excluded during testing, because they are known to be negative due to their incompatibility. Used in this way, incompatibility is employed as a hard constraint on coreference (e.g. by Strube & Müller (2003)). Alternatively, incompatibility can be left to the classifier to detect. This has the advantage that potential exceptions can be accomodated in the model that is learned by the classifier. Used in this way, the relations between number, gender, and person of anaphor and antecedent are employed as normal features.

In the feature set used in this thesis, incompatibility in the categories number, gender and person is represented by the features **NumberRelation**, **GenderRelation**, and **PersonRelation** (cf. Chapter 6.2.2), respectively. For training and testing in our experiments, we used this information either as a hard constraint or as a normal feature. In the former case, an instance was dropped if it had the value *incompatible* for any of the three features mentioned above. In the latter case, incompatible instances were not dropped but allowed in the data set. As a rule, the same instance filtering method was used for training and testing: If incompatible instances were removed during training, they were also removed for testing, and vice versa.

7.2.5 Corpus-based Features

Chapters 6.2.1 and 6.2.2 introduced a couple of quantitative features which are based on counts in a huge text corpus. Some of these features capture semantic compatibility between e.g. an NP antecedent and the predicative context of the anaphor. Features

[49] Schiffman (1985, p.59) even claims the construction *do it* to be a positional alternant of *it* if the antecedent is a VP.

of this type are already rather common in anaphora resolution (e.g. Yang et al. (2003)). However, there is no apparent consensus about whether these features are really helpful (see e.g. Kehler et al. (2004) and Bean & Riloff (2004)), so it seems interesting to evaluate their contribution for the current task.

Other corpus-based features try to capture e.g. an anaphor's preference for a VP rather than an NP antecedent. Most of these features operationalize observations made in the linguistic literature (e.g. Eckert & Strube (2000)) and have not been used in this manner before, and again we are interested in their potential contribution.

For this reason, we run different experimental setups which do or do not use the corpus-based features. Rather than dealing with all features separately, we chunk them together to a group of corpus-based features whose effect is studied as a whole. If the features are used, they are available during both training and testing. Likewise, if they are not used, they are not available in either phase.

7.2.6 Processing Order: Discourse vs. Chronologial Order

This parameter relates to the question in which order the expressions in the dialog are processed: *discourse order* is the order in which the expressions appear in the transcript, while *chronological order* is the order defined by the expressions' time-stamps. None of the spoken dialog anaphora resolution approaches that we are aware of does make this distinction. Rather, processing is always done in discourse order, apparently under the assumption that discourse order is sufficiently similar to chronological order. This assumption is probably true for two-party dialog. However, in the context of multi-party dialog with considerable overlap, the sequential ordering of the utterances in the transcript does not adequately reflect the actual sequence. As was described in Chapter 3.1, the segmentation of the ICSI Meeting Corpus was deliberately simplified in order to prevent a high degree of fragmentation. The resulting discourse order can thus be expected to be sometimes arbitrary. Consequentially, it seems reasonable to refer to the chronological ordering in order to capture more adequately the sequence of utterances. What is more, in a realistic setting where speech recognizer output from several channels is used instead of manually segmented transcripts, discourse order is not available. We performed some initial resolution experiments using discourse vs. chronological order. Interestingly, these experiments showed chronological order to generally yield worse resolution results than discourse order. Since the chronological order that we use is potentially inaccurate because it was created on the basis of a simple forced alignment (see Chapter 6.1.3), it is unclear whether the worse performance is due to the chronological order *per se*, or due to the fact that the alignment is faulty. Therefore, in order to keep the amount of data (and thus the number of result tables) low, we only report results obtained using discourse order.

7.2.7 Type of Classifier

The machine learning classifier is the crucial component for locally deciding whether the anaphor in the current anaphor-antecedent pair actually stands in an anaphoric relation to the antecedent. It is this component that is built resp. parameterized during the training phase of machine learning. All classifiers used in this thesis come from the WEKA machine learning workbench (Witten & Frank, 1999). Training and test instances are commonly represented as feature vectors, i.e. lists of attribute-value pairs. For WEKA, attributes, their types (numeric or nominal) and their possible values (for nominal attributes) need to be predefined in a special section of the file containing the instances. The actual instances are then only specified as an ordered list of *values*, where each list position is associated with a particular attribute. The complete list of features used in our experiments has been described in Chapter 6.2. In the following, we use the schematic notation I = (f_1,f_2,f_3,...,f_n) to represent a single data instance. Each instance is assigned a class label which is either ′true′ (positive instance) or ′false′ (negative instance). Positive instances are those in which the anaphor actually stands in an anaphoric relation to the antecedent, while the rest are negative instances.

For the present work, we chose a range of classifiers exemplifying different machine learning paradigms. Unless noted otherwise, they are used in their default settings. The following brief description of these classifiers is only intended to provide a rough, non-technical characterization, outlining their major similarities and differences. For more detailed information, see e.g. Mitchell (1997), Bishop (2006), and the references provided below.

A basic distinction can be made between so-called *lazy* and *eager* learners (Mitchell, 1997). For lazy (or *instance-based*) learners, training amounts to simply storing all provided data instances. Classification of an instance with unknown class is done by comparing it to the stored instances for which the classes are known, and choosing a class on the basis of similarity. A well-known instance-based learner which has also been applied to anaphora resolution by e.g. Preiss (2002) and Hoste (2005) is TiMBL (Daelemans et al., 2004). In our experiments, we used the k-nearest-neighbour classifier IBk from WEKA, with the number of nearest neighbours ($= k$) set to 2.

The majority of machine learning systems belong to the class of eager learners. These learners do not store individual data instances during training, but they build generalized models. The form of these models varies: Some learners produce decision trees or sets of rules, which has the additional advantage that these models can be inspected and interpreted by humans. In our experiments, the J48 classifier belongs to this category. The J48 classifier is a Java version of the well-known C4.5 decision tree learner (Quinlan, 1993). C4.5 resp. J48 have commonly been used for coreference resolution, especially in systems employing the mention-pair approach (McCarthy & Lehnert, 1995; Aone & Bennett, 1995; Ng & Cardie, 2002; Müller et al., 2002). The J48 classifier builds a tree-like structure by recursively partitioning the set of training instances. This partitioning depends on a particular attribute that is tested at each tree node. The crucial task that the decision tree learner has to solve is to choose the *optimal* attribute to be tested at a given tree node. For C4.5 resp. J48, this choice is made on the basis of the *information gain* of the attribute. The information gain quantifies how well an attribute separates instances in the training data set into subsets of different classes. For each node, the learner greedily chooses the attribute with the highest information gain. As Mitchell (1997) points out, this greediness leads to trees in which the most informative attributes tend to be close to the top.

Other learners produce models in the form of mathematical equation systems. From this category, we employ the Naive Bayes and the Logistic Regression classifier. The Naive Bayes classifier builds a model by calculating for each feature in I = (f_1,f_2,f_3,...,f_n) and for each class 'true' and 'false' the conditional probability of that feature given that class. The conditional probability is simply estimated as the relative frequency in the training data, i.e. as the frequency of a feature f_n appearing in a training instance of class c_i, divided by the total frequency of f_n. For classification of an unseen instance I = (f_1,f_2,f_3,...,f_n), the model is employed for choosing that class label c_i for which the product of the conditional probabilities of all features in I is maximal.

Since it calculates the conditional probability of an instance having a particular class by simply multiplying the individual conditional probabilities, the Naive Bayes classifier assumes that the individual features are statistically independent of each other. Although this independence assumption in practice is often false, Naive Bayes has been shown to perform surprisingly well (Rish, 2001; Rish et al., 2001). Ng & Cardie (2003) have employed it for coreference resolution (as one of two classifiers in a co-training ensemble). They point out that one of the advantages of Naive Bayes classifiers which makes them particularly appropriate for coreference resolution is that they tend to be robust against unbalanced data sets (cf. Chapter 7.2.2 above).

The last of the four classifiers used in our experiments is Logistic Regression[50] (Agresti, 1990; Hosmer & Lemeshow, 2000). In the past, Logistic Regression has mainly been employed as a descriptive tool for modelling if and to what degree the features of an instance are able to explain the instance's class label. One of the few applications in computational linguistics is Strube & Wolters (2000). Their work deals with a probabilistic model of pronominalization, and Logistic Regression is used to describe the explanatory power of features like agreement, syntactic function, and parallelism for predicting whether or not an anaphor is realized as a proform or as a full noun phrase.

For Logistic Regression, training amounts to the estimation of a number of parameters. The principles underlying Logistic Regression can best be described in analogy to the simpler linear regression. In linear regression, the goal is to find a function which describes as accurately as possible the relation between two numerical variables (viewed as the function *argument* and the function *value*). The data to be modelled is given as a series of observed argument-value pairs ($< x_1, y_1 >$, $< x_2, y_2 >$, ..., $< x_n, y_n >$). In principle, there can be a huge number of functions which approximate these argument-value pairs. A common criterion for selecting a function is to choose the one which minimizes the sum of the squared differences between the observed values and those returned by the function. Once a function has been estimated in this way, it can be used to calculate, for a given x_i, the pertaining y_i.

The main difference of Logistic Regression is that the value of the function that is to be estimated is not numeric but binary,[51] i.e. in our case 'true' or 'false'. Since the function value is not numeric, there is no numeric difference between the observed value and the value predicted by the function. Therefore, the method of minimizing the squared differences between both is not applicable. Instead, the function for describing the relation between the function arguments (i.e. each feature of I = (f_1,f_2,f_3,...,f_n)) and the binary

[50] Simply called *Logistic* in WEKA.

[51] This holds for the 'standard' Logistic Regression. The *Logistic* classifier in WEKA is an extension which supports multinomial classification, but since we use it for binary classification, these extensions have no effect.

function value is estimated using a *maximum likelihood* approach (Manning & Schütze, 1999). In this approach, the probability of each of the two possible values 'true' and 'false' is modelled as a function of the Logistic Regression model parameters to be estimated. These unknown parameters are chosen in such a way that they fit as exactly as possible the actual distribution observed in the training data.
Just like Naive Bayes, Logistic Regression makes a feature independence assumption, i.e. it also calculates the overall probability of an instance I = (f_1,f_2,f_3,...,f_n) by multiplying the individual probabilities.

7.2.8 Filtering Non-Referential *It*

As was described in Chapter 4.1.2, 37.5% of all instances of *it* in our five dialogs are non-referential. Ideally, these instances have to be prevented from entering into an anaphoric relation, because they probably introduce spurious links and thus precision errors. Chapter 6.1.1 described a component for automatic detection of non-referential *it*. When evaluated in isolation, the performance of this component is precision 80.0, recall 60.9, and F-measure 69.2. While especially the precision is fairly good, it is still not perfect, and there is a considerable risk that the component also harms recall by falsely removing instances of *it* from consideration as anaphors or antecedents. Therefore, we performed one set of experiments in which the automatic filter component was used during testing, and another set in which it was not used during testing.[52] During training, on the other hand, the filter component was always used.

7.2.9 Resolution Algorithm

The resolution algorithm provides the central processing framework for the interaction of all resolution parameters described so far. As was already mentioned, we deliberately chose a simple and well-understood algorithm, implementing the mention-pair approach. With this approach we model the resolution of *it*, *this* and *that* in spoken dialog as a binary classification task, i.e. as the mapping of anaphoric mentions to previous mentions of the same referent. This has been the standard in computational pronoun resolution for quite some time. The algorithm is locally limited and cannot take into account information beyond that which is conveyed by the two expressions. There are some alternative approaches which model pronoun resolution *incrementally* by taking

[52] See Chapter 7.5 for a description of experiments with manually improved data.

earlier resolution decisions into account (e.g. the direct antecedent (if any) of the current potential antecedent (Iida et al., 2003), or the entire chain of the antecedent so far (Luo et al., 2004) (see also Chapter 5.1). Since approx. 75% of the anaphoric chains in our data set would not benefit from incremental processing because they contain one anaphor only (see Chapter 4.2.5.4), we think that the limitation brought about by using a local model is not serious. In addition, incremental processing bears the risk of system degradation due to error propagation.

The resolution algorithm is outlined in Algorithm 6. The resolution algorithm mainly consists of one forward iteration over all NP and VP chunks in discourse order (6-44).[53] Within this iteration, only potential anaphors, i.e. chunks matching *it*, *this*, or *that*, are considered (8). Potentially anaphoric instances of *it* that are automatically identified as non-referential are skipped (10), if the respective option is used. For all other potential anaphors, the algorithm tries to identify a correct antecedent. The antecedent search is implemented as a backward iteration over all chunks (14-39). Just like the potential anaphors, potential antecedents are also filtered for instances of *it* that are automatically identified as non-referential (17), if the respective option is used. Instances in which anaphor and antecedent would be more than the respective maximal search depth apart are dropped (20 resp. 23). Instances with VP antecedents are also dropped unless the anaphor is *that* or unless the anaphor is the object in a construction involving a form of the verb *do* (25). Then, further plausibility criteria are checked (29). If prefiltering of incompatible instances is performed (see Chapter 7.2.4), this is also implemented by means of plausibility checks. If the pair of anaphor and antecedent is not filtered out by the plausibility checks, it is submitted to the classifier for classification (32). If it is classified as 'true', i.e. as actually being coreferent, the classifier's confidence is determined (33). If this confidence is higher than any other confidence previously stored for the current potential anaphor, both the new maximum confidence and the potential antecedent that produced it are stored (34-37). After all potential antecedents have been tested with the current potential anaphor, the algorithm selects as the most probably correct one the one (if any) which yielded the highest confidence for 'true' when paired with the current anaphor. The resolution result is stored by adding a link between anaphor and antecedent to the set of *response* links (41).

[53] In the following (Algorithm 6 and Algorithm 7), numbers refer to line numbers.

Algorithm 6 Algorithm `ResolveMentionPair`

```
chunks $\leftarrow$ all NP and VP chunks in discourse order
maxNPDist $\leftarrow$ (13|9|6)
maxVPDist $\leftarrow$ (7|7|3)
classifier $\leftarrow$ (Baseline|J48|NaiveBayes|LogisticRegression|IB2)
NONREF-IT-FILTER $\leftarrow$ (true|false)
for $i = 1$ to chunks.size() do
    currentAna $\leftarrow$ chunks.get(i)
    if currentAna equals 'it','this','that' then
        if NONREF-IT-FILTER = true and onAutoNonRefLevel(currentAna) then
            continue
        end if
        maxConf $\leftarrow 0.0$
        maxConfAnte $\leftarrow null$
        for $j = i - 1$ to 0 do
            currentAnte $\leftarrow$ chunks.get(j)
            if NONREF-IT-FILTER = true and onAutoNonRefLevel(currentAnte) then
                continue
            end if
            if currentAnte.type() = NP and dist(currentAnte,currentAna) $>$ maxNPDist then
                continue
            else if currentAnte.type() = VP then
                if dist(currentAnte,currentAna) $>$ maxVPDist then
                    continue
                end if
                if currentAna equals 'that' = false and do-Object(currentAna) = false then
                    continue
                end if
            end if
            if plausible(currentAnte,currentAna) = false then
                continue
            end if
            if classify(currentAnte,currentAna,classifier) equals 'true' then
                currentConf $\leftarrow$ getConf(currentAnte,currentAna,classifier)
                if currentConf $>$ maxConf then
                    maxConf $\leftarrow$ currentConf
                    maxConfAnte $\leftarrow$ currentAnte
                end if
            end if
        end for
        if maxConfAnte != null then
            addLinkToResponse(currentAna,maxConfAnte)
        end if
    end if
end for
```

7.3 Training Data Generation

The resolution experiments in this thesis are performed by means of dialog-wise cross-validation. For each cross-validation run, four of our five dialogs were used as training data, and the remaining fifth as test data. The resolution algorithm was already described in Chapter 7.2.9. Here, we describe the algorithm for training data generation.

The algorithm is outlined in Algorithm 7. The training data generation algorithm mainly consists of one forward iteration over all NP and VP chunks in discourse order (5-67). Within this iteration, only potential anaphors, i.e. chunks matching *it*, *this*, or *that*, are considered (7). If the current chunk is a potential anaphor, its *corefClass* attribute is evaluated.[54] If the current potential anaphor is not part in any coreference set (*corefClass* = 'none'), this can have two reasons.

The first reason is that the potential anaphor is referential, but that it is missing on the current training data level (core level 2, 3, or 4) because it failed to meet the respective criteria of being identified by at least two, three, or four annotators, respectively. In this case, neither positive nor negative instances can be created from the potential anaphor, because no information is available about its true resp. false antecedents.

The second reason for the potential anaphor not being part in any coreference set is that the potential anaphor is non-referential (i.e. pleonastic or *prop*-it resp. discarded). In this case, negative instances are created from the potential anaphor and all of its antecedents within the maximal search depth. This is intended to make the classifier more robust for handling non-referential instances of *it*, *this*, and *that*. In the training data generation algorithm, a potential anaphor is identified as non-referential by checking if there is a corresponding markable on the manually created markable level *manualNonRef* (13). This level contains markables for all instances of *it*, *this*, and *that* that were annotated as non-referential by at least three out of four annotators. If the current anaphor is found to be non-referential, training data instances are created from it by pairing it with all preceeding chunks as potential antecedents. This is accomplished in a backward iteration over all chunks (14-33). Instances in which anaphor and antecedent would be more than the respective maximal search depth apart are dropped (20 resp. 23). Instances with VP antecedents are also dropped unless the anaphor is *that* or unless the anaphor is the object in a construction involving a form of the verb *do* (25). Then, further plausi-

[54] As was described in Chapter 4.2.2, the *corefClass* attribute encodes the membership of a chunk resp. a markable in a coreference set. Chunks which have been annotated as belonging to the same class have the same value in this attribute. If a chunk is not assigned to any coreference class, it has the value 'none'.

bility criteria are checked (29). If prefiltering of incompatible instances is performed (see Chapter 7.2.4), this is also implemented by means of plausibility checks. If the pair of anaphor and antecedent is not filtered out by the plausibility checks, a negative instance is created from it and added to the training data basis.

If the current potential anaphor is found to be a member in a coreference set (35), one positive as well as several negative instances have probably to be created for it. Again, pairs of anaphors and antecedents are created by means of a backward iteration over all chunks (37-64). Distance and plausibility checking is performed as described above. If the *corefClass* attribute of the potential anaphor does not match that of the current potential antecedent, the pair gives rise to a negative instance. Rather than creating and adding the negative instance immediately, the algorithm only stores the antecedent part of it (56). The reason is that we want to create negative instances for a potential anaphor only if the anaphor also gives rise to a positive one. A potential anaphor can fail to give rise to a positive instance in spite of being referential e.g. if the antecedent lies outside the maximal search depth. Therefore, creation and adding of the negative instance(s) is postponed until a positive instance is created as well.

If the *corefClass* attribute of the potential anaphor matches that of the potential antecedent, the pair gives rise to a positive instance. This instance is added to the training data basis (58), along with all negative instances resulting from the pairing of the current potential anaphor and all incorrect antecedents encountered prior to the correct one (59-61). After a positive instance has been added for the current potential anaphor, the search for further antecedents is terminated (60). Thus, negative instances are created only for antecedents occurring before the antecedent which gives rise to the positive instance. This practice of limiting the amount of negative instances is also commonly employed, e.g. by Soon et al. (2001) and Strube & Müller (2003).

Algorithm 7 Algorithm `GenerateMentionPair`

```
chunks ← all NP and VP chunks in discourse order (from core data set 2,3,4)
maxNPDist ← (13|9|6)
maxVPDist ← (7|7|3)
NONREF-IT-FILTER ← true
for $i = 1$ to chunks.size() do
    currentAna ← chunks.get(i)
    if currentAna equals 'it','this','that' then
        anaCorefClass ← currentAna.getCorefClass()
        if anaCorefClass equals 'none' then
            if NONREF-IT-FILTER = true and onAutoNonRefLevel(currentAna) then
                continue
            end if
            if onManualNonRefLevel(currentAna) then
                for $j = i - 1$ to 0 do
                    currentAnte ← chunks.get(j)
                    if NONREF-IT-FILTER = true and onAutoNonRefLevel(currentAnte) then
                        continue
                    end if
                    if currentAnte.type() = NP and dist(currentAnte,currentAna) $>$ maxNPDist then
                        continue
                    else if currentAnte.type() = VP then
                        if dist(currentAnte,currentAna) $>$ maxVPDist then
                            continue
                        end if
                        if currentAna equals 'that' = false and do-Object(currentAna) = false then
                            continue
                        end if
                    end if
                    if plausible(currentAnte,currentAna) = false then
                        continue
                    end if
                    addTrainingInstance(currentAna,currentAnte,'false')
                end for
            end if
        else
            negativeList.empty()
            for $j = i - 1$ to 0 do
                currentAnte ← chunks.get(j)
                if NONREF-IT-FILTER = true and onAutoNonRefLevel(currentAnte) then
                    continue
                end if
                if currentAnte.type() = NP and dist(currentAnte,currentAna) $>$ maxNPDist then
                    continue
                else if currentAnte.type() = VP then
                    if dist(currentAnte,currentAna) $>$ maxVPDist then
                        continue
                    end if
                    if currentAna equals 'that' = false and do-Object(currentAna) = false then
                        continue
                    end if
                end if
                if plausible(currentAnte,currentAna) = false then
                    continue
                end if
                if anaCorefClass equals currentAnte.getCorefClass() = false then
                    negativeList.add(currentAnte)
                else
                    addTrainingInstance(currentAna,currentAnte,'true')
                    for all Chunks currentNegativeAnte in negativeList do
                        addTrainingInstance(currentAna,currentNegativeAnte,'false')
                    end for
                    break
                end if
            end for
        end if
    end if
end for
```

The following Table 43 gives the number of positive and negative training data instances produced by all five dialogs for all settings. The setting +/-filter controls whether number, gender, and person incompatibility are treated as constraints (+filter) or as features (-filter) (Chapter 7.2.4). The percentage of negative instances displayed in column **Neg %** is the resulting negative bias of the training data set. With increasing oversampling rate, the bias decreases: For 3-fold oversampling, the negative bias is in the high 60s to low 70s, while for 6-fold oversampling, it is in the high 40s to low 50s. 6-fold oversampling thus produces a data set that is almost balanced.

Setting		**2**			**3**			**4**		
		Pos	**Neg**	**Neg %**	**Pos**	**Neg**	**Neg %**	**Pos**	**Neg**	**Neg %**
OS rate 0	-filter	379	2364	86.18%	246	1549	86.30%	101	815	88.97%
	+filter	350	1992	85.06%	226	1295	85.14%	94	682	87.89%
OS rate 3	-filter	1137	2364	67.52%	738	1549	67.73%	303	815	72.90%
	+filter	1050	1992	65.48%	678	1295	65.64%	282	682	70.75%
OS rate 6	-filter	2274	2364	50.97%	1476	1549	51.21%	606	815	57.35%
	+filter	2100	1992	48.68%	1356	1295	48.85%	564	682	54.74%

Table 43: Training data sets.

The table also nicely shows the effect of the different parameter settings. Using incompatibility as a hard constraint (+filter) on instance generation consistently causes fewer positive and fewer negative instances to be created. The decrease in negative instances is consistently higher than that in positive instances, leading to a weakening of the negative bias in all training data sets. In view of the fact that positive instances are rare, however, it seems doubtful if sacrificing positive instances in favour of a weaker negative bias is reasonable.

7.4 Resolution Experiments with Automatically Obtained Data

This chapter presents the results of resolution experiments that were performed in a realistic, real-world setting. This means that for these experiments, only fully automatically obtained information about disfluencies and non-referential *it* was employed. The experimentation included the systematic variation of some of the experimental parameters described above, and the determination of the setting which yielded the best overall result. The performance of the best-performing system is the final result for the fully automatic resolution. This system is also be the basis for exploring potential improvement with the help of idealized input in Chapter 7.5.

The parameters whose settings were systematically varied are

- **Training data set (core data set 2, 3, or 4)**
- **Rate of oversampling (0, 3, or 6)**
- **Instance pre-filtering (+/-filter)**
- **Non-Referential *it* filtering (+/-it-filter)**
- **Corpus-based features (+/-tipster)**

We report precision (P), recall (R) and F-measure (F) based on the definition in Chapter 7.1. Note that all results were obtained using exactly the same key data set, i.e. core data set 3, so that performance differences are entirely due to differences in the testing setup. Throughout the presentation of results, the maximum F-measure yielded for every classifier (2-Trained, 3-Trained, 4-Trained, cf. below) and for every oversampling rate (0, 3, 6) is highlighted. Separate result figures are reported for overall performance (ALL) and for NP and VP antecedents alone. For each table and for each classifier (2-, 3-, or 4-Trained), the highest *overall* F-measure is highlighted in grey. In addition, the respective highest F-measures for NP and VP antecedents are highlighted in a lighter grey.

Baseline We implemented a simple recency-based algorithm as a reference baseline. The baseline system simply resolved instances of *it*, *this* and *that* to the most recent matching[55] antecedent. In analogy to using three classifiers trained on different data sets, for the baseline algorithm we performed three runs with different maximal antecedent search depths (13, 9, and 6 seconds for NP antecedents and 7 and 3 seconds for VP antecedents). The features +tipster and -filter do not apply for the baseline algorithm: Tipster features are available for learning-based (i.e. non-baseline) classifiers only, and instance filtering is always performed in the baseline algorithm *qua* selection of the most recent compatible antecedent.

The baseline algorithm's recall for NP antecedents is considerable, ranging slightly above 26. Recall for NP antecedents was bound to suffer with the inclusion of VP antecedents into resolution, because the baseline algorithm will indiscriminately treat as

[55] An antecedent matches an anaphor if they are not *incompatible* in either **NumberRelation**, **GenderRelation**, or **PersonRelation**, and if **ArgumentOf** (for VP antecedents) and **CoArguments** (for NP antecedents) are both *false* (cf. Chapter 6.2).

					13/7			9/7			6/3	
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-
		+tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-
	+it-filter	-tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-
		+tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-
+filter	-it-filter	-tipster	NP	4.67	27.97	8.01	4.62	27.12	7.90	4.74	26.27	8.03
			VP	1.71	2.63	2.07	1.72	2.63	2.08	1.98	2.63	2.26
			ALL	4.45	21.32	7.36	4.40	20.69	7.25	4.54	20.06	7.40
		+tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-
	+it-filter	-tipster	NP	5.01	25.85	**8.40**	5.18	26.27	**8.65**	5.38	25.85	**8.91**
			VP	1.75	2.63	**2.11**	1.77	2.63	**2.12**	2.04	2.63	**2.30**
			ALL	4.73	19.75	**7.64**	4.88	20.06	**7.85**	5.12	19.75	**8.13**
		+tipster	NP	-	-	-	-	-	-	-	-	-
			VP	-	-	-	-	-	-	-	-	-
			ALL	-	-	-	-	-	-	-	-	-

Table 44: Baseline results.

potentially discourse-deictic all pronouns that are either *that* or that are the object of a form of *do*. This means that many instances of *that* will be wrongly resolved to VP antecedents just because a VP antecedent is encountered before any NP antecedent. For the same reason, F-measure for VP antecedents is extremely low for the baseline algorithm (2.30). It is striking that the recall for VP antecedents is exactly the same for all six baseline algorithm settings (2.63). The recall for ALL is consistently about 20, so the baseline algorithm is able to find the correct NP and VP antecedent for roughly 1/5 of all anaphors. The best F-measure is 8.13, yielded by the setting +it-filter using the minimal antecedent search depth. Thus, the baseline algorithm yields the best results in the most restricted setting. This is not surprising, since the main problem of the baseline algorithm is its low precision, which can best be countered by maximally restricting its greediness.

J48 Unlike the baseline algorithm, the J48 classifier (as well as all other non-baseline systems in the following) is built from a training data set. For training, several different training data sets (cf. Chapter 7.3) were available, which differed with respect to the core data sets (2, 3, or 4) and with respect to the rate of positive oversampling (none, 3-fold, 6-fold). In the following, we present one table of results for each of the three oversampling rates. Within each table, we provide three sets of results for each different testing setup (+/-tipster, +/-it-filter, +/-filter). These sets of results differ with respect to how *constrained* the classifier is, i.e. with respect to the nature of the core data set and the maximal antecedent search depth that was used for training. For example, the column **2-Trained** contains the results obtained by using the classifier that was trained on the rather large and unconstrained core data set 2 with the maximal antecedent search depths determined for that data set, i.e. 13 seconds for NP and 7 seconds for VP. The column **4-Trained**, on the other hand, contains results obtained by using the classifier that was trained on the much smaller and more strictly defined core data set 4 with a maximal NP search depth of 6 seconds and VP search depth of 3. As a rule, the same maximal antecedent search depth that was used during training was also used during testing.

A striking fact about the non-oversampled J48 classifier in Table 45 is that the 4-Trained version fails to find even a single VP antecedent. The best F-measure for ALL is 11.45, yielded by the 2-Trained classifier using the setting +tipster, +it-filter, +filter. This setting also produces the highest F-measure for NP antecedents, 13.33.

				2-Trained			**3-Trained**			**4-Trained**		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	8.53	7.63	8.05	13.36	12.29	12.80	10.17	2.54	4.07
			VP	3.28	5.26	4.04	4.40	5.26	4.79	0.00	0.00	0.00
			ALL	6.89	7.21	7.04	10.71	10.35	10.53	10.17	1.88	3.18
		+tipster	NP	5.08	9.32	6.58	11.68	14.41	**12.90**	8.93	2.12	3.43
			VP	3.57	5.26	4.26	3.30	3.95	3.59	0.00	0.00	0.00
			ALL	4.77	8.15	6.02	9.69	11.60	10.56	6.76	1.57	2.55
	+it-filter	-tipster	NP	9.28	7.63	8.37	12.90	10.17	11.37	9.62	2.12	3.47
			VP	3.28	5.26	4.04	4.30	5.26	4.73	0.00	0.00	0.00
			ALL	7.26	7.21	7.23	10.04	8.78	9.37	9.62	1.57	2.70
		+tipster	NP	4.81	7.63	5.90	11.20	11.86	11.52	9.80	2.12	3.48
			VP	3.45	5.26	4.17	3.26	3.95	3.57	0.00	0.00	0.00
			ALL	4.49	6.90	5.44	9.06	9.72	9.38	7.25	1.57	2.58
+filter	-it-filter	-tipster	NP	11.06	10.59	10.83	8.49	13.14	10.32	17.24	2.12	3.77
			VP	2.63	3.95	3.16	6.45	5.26	5.80	0.00	0.00	0.00
			ALL	8.50	9.09	8.79	8.20	10.97	9.38	17.24	1.57	2.87
		+tipster	NP	11.71	14.83	13.08	10.26	16.53	12.66	18.18	3.39	5.71
			VP	5.22	7.90	**6.28**	6.45	5.26	5.80	0.00	0.00	0.00
			ALL	10.12	13.17	11.44	9.73	13.48	**11.30**	12.90	2.51	4.20
	+it-filter	-tipster	NP	11.06	9.32	10.12	7.94	11.44	9.38	20.00	2.12	3.83
			VP	2.63	3.95	3.16	6.45	5.26	5.80	0.00	0.00	0.00
			ALL	8.28	8.15	8.22	7.71	9.72	8.60	20.00	1.57	2.91
		+tipster	NP	12.74	13.98	**13.33**	10.11	15.25	12.16	20.93	3.81	**6.45**
			VP	5.00	7.90	6.12	6.67	5.26	**5.88**	0.00	0.00	0.00
			ALL	10.53	12.54	**11.45**	9.62	12.54	10.88	14.75	2.82	**4.74**

Table 45: J48 results.

				2-Trained			3-Trained			4-Trained		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	6.85	24.58	10.71	7.90	22.03	11.63	7.04	11.86	8.83
			VP	4.69	7.90	5.88	3.54	5.26	4.23	7.81	6.58	7.14
			ALL	6.54	20.06	9.87	7.49	18.18	10.61	7.31	10.66	8.67
		+tipster	NP	6.44	23.73	10.13	7.24	19.92	10.62	6.44	11.44	8.24
			VP	10.00	15.79	**12.25**	3.60	5.26	4.28	2.04	1.32	1.60
			ALL	6.85	21.32	10.37	6.81	16.30	9.60	6.17	9.09	7.35
	+it-filter	-tipster	NP	6.64	21.19	10.11	7.65	19.07	10.92	6.54	10.17	7.96
			VP	4.69	7.90	5.88	3.54	5.26	4.23	8.33	6.58	**7.35**
			ALL	6.34	17.56	9.31	7.10	15.67	9.78	6.98	9.40	8.01
		+tipster	NP	7.09	23.31	10.87	7.75	19.07	11.02	6.84	11.02	8.44
			VP	9.84	15.79	12.12	3.51	5.26	4.21	2.13	1.32	1.63
			ALL	7.44	21.00	**10.98**	7.15	15.67	9.82	6.53	8.78	7.49
+filter	-it-filter	-tipster	NP	7.33	23.31	11.16	8.07	20.34	11.55	8.00	11.86	**9.56**
			VP	7.84	15.79	10.48	6.42	9.21	7.57	4.84	3.95	4.35
			ALL	7.41	21.00	10.96	7.91	17.56	10.91	7.97	10.35	**9.00**
		+tipster	NP	7.33	24.15	**11.24**	9.24	22.03	**13.02**	7.79	10.59	8.98
			VP	6.57	11.84	8.45	4.92	7.90	6.06	1.75	1.32	1.50
			ALL	7.19	20.69	10.67	8.56	18.50	11.71	6.88	8.15	7.46
	+it-filter	-tipster	NP	7.30	20.76	10.81	8.58	19.49	11.92	7.49	9.75	8.47
			VP	7.74	15.79	10.39	6.96	10.53	**8.38**	5.09	3.95	4.44
			ALL	7.38	19.12	10.65	8.40	17.24	11.29	7.34	8.46	7.86
		+tipster	NP	6.98	20.34	10.39	9.52	20.34	12.97	6.36	7.63	6.94
			VP	7.52	13.16	9.57	6.35	10.53	7.92	1.79	1.32	1.52
			ALL	7.04	18.18	10.15	8.99	17.87	**11.96**	5.61	5.96	5.78

Table 46: J48 results, 3-fold oversampling.

The effect of 3-fold oversampling on the J48 classifier can be observed in Table 46. Oversampling has a positive effect on F-measure for ALL in 20 out of 24 settings. For NP, the rate is 18 out of 24, and for VP, it is even 22 out of 24. The latter effect is mainly due to the fact that 3-fold oversampling causes the 4-Trained classifier to now find VP antecedents in all eight settings where it failed to find any before. 3-fold oversampling brings about a slightly improved best overall F-measure of 11.96, produced by the 3-Trained classifier in the setting +tipster, +it-filter, +filter.

				2-Trained			3-Trained			4-Trained		
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	6.44	23.73	10.13	9.35	26.70	13.85	6.61	12.29	8.59
			VP	4.17	6.58	5.10	3.16	3.95	3.51	5.33	5.26	**5.30**
			ALL	6.16	19.12	9.31	8.69	21.00	12.29	6.41	10.35	7.91
		+tipster	NP	6.91	25.42	10.87	6.95	20.34	10.36	5.88	10.59	7.56
			VP	8.16	10.53	**9.20**	1.22	1.32	1.27	0.00	0.00	0.00
			ALL	7.03	21.32	10.57	6.44	15.67	9.13	5.01	7.84	6.11
	+it-filter	-tipster	NP	6.50	21.61	9.99	10.12	25.42	**14.48**	6.12	10.17	7.64
			VP	5.65	9.21	7.00	2.02	2.63	2.29	5.33	5.26	**5.30**
			ALL	6.37	18.18	9.44	9.22	20.06	**12.64**	5.98	8.78	7.12
		+tipster	NP	6.96	22.88	10.67	7.15	18.64	10.34	6.04	9.75	7.46
			VP	7.41	10.53	8.70	1.19	1.32	1.25	0.00	0.00	0.00
			ALL	7.00	19.44	10.29	6.55	14.42	9.01	5.06	7.21	5.94
+filter	-it-filter	-tipster	NP	7.07	22.88	10.80	6.97	18.22	10.08	7.35	10.59	8.68
			VP	3.13	5.26	3.92	2.94	3.95	3.37	4.17	3.95	4.05
			ALL	6.47	18.18	9.54	6.66	15.05	9.23	6.78	8.78	7.65
		+tipster	NP	7.87	25.42	**12.02**	7.72	20.34	11.19	8.61	12.29	**10.12**
			VP	5.65	9.21	7.00	3.64	5.26	4.30	0.00	0.00	0.00
			ALL	7.54	21.00	**11.09**	7.10	16.30	9.90	7.13	9.09	**7.99**
	+it-filter	-tipster	NP	7.01	20.76	10.53	6.79	16.10	9.55	6.76	8.48	7.52
			VP	3.08	5.26	3.88	3.00	3.95	3.41	4.11	3.95	4.03
			ALL	6.39	16.61	9.23	6.34	13.17	8.56	6.22	7.21	6.68
		+tipster	NP	7.76	22.88	11.59	8.10	19.49	11.44	8.22	10.59	9.26
			VP	4.17	6.58	5.10	3.77	5.26	**4.40**	0.00	0.00	0.00
			ALL	7.20	18.50	10.37	7.42	15.67	10.07	6.65	7.84	7.19

Table 47: J48 results, 6-fold oversampling.

6-fold oversampling leads to an overall decrease in performance except for a few settings. Notably, the 4-Trained classifier again fails to find VP antecedents in all four settings involving the feature +tipster. Despite the overall downwards trend in performance, the 3-Trained classifier trained on the 6-fold oversampled data yields a new best overall F-measure of 12.64. It is accompanied by another best NP F-measure of 14.48.

For the J48 classifier, the best F-measure for ALL is produced by using the 3-Trained classifier (6-fold oversampling) in the setting -tipster, +it-filter, and -filter. Precision is 9.22, recall is 20.06, and F-measure 12.64. This classifier also yields the best NP F-measure of 14.48. VP performance, however, is very poor with an F-measure of only 2.29.

Naive Bayes The WEKA implementation of the Naive Bayes classifier has the option of using supervised discretization for numerical attributes. Using this feature yielded consistently better results, so we only report results produced with this option active.

				2-Trained			**3-Trained**			**4-Trained**		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	11.25	30.09	16.38	9.71	31.36	14.83	11.83	19.49	14.72
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.69	22.26	14.45	9.71	23.20	13.69	11.83	14.42	12.99
		+tipster	NP	11.39	30.51	16.59	9.53	30.93	14.57	11.11	18.22	13.80
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.80	22.57	14.60	9.53	22.88	13.46	11.11	13.48	12.18
	+it-filter	-tipster	NP	13.30	32.63	**18.90**	10.71	30.51	15.86	12.02	17.37	14.21
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	12.64	24.14	**16.60**	10.71	22.57	14.53	12.02	12.85	12.42
		+tipster	NP	12.80	31.36	18.18	10.50	30.09	15.57	12.09	17.37	14.26
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	12.13	23.20	15.93	10.50	22.26	14.27	12.09	12.85	12.46
+filter	-it-filter	-tipster	NP	10.97	29.66	16.02	11.04	34.75	16.75	13.30	20.34	16.08
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.90	21.94	14.57	11.04	25.71	15.44	13.30	15.05	14.12
		+tipster	NP	10.85	29.24	15.83	11.04	34.75	16.75	12.71	19.49	15.39
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.76	21.63	14.38	11.04	25.71	15.44	12.71	14.42	13.51
	+it-filter	-tipster	NP	12.00	29.24	17.02	11.92	32.63	**17.46**	14.29	19.07	**16.33**
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	11.90	21.63	15.35	11.92	24.14	**15.96**	14.29	14.11	**14.20**
		+tipster	NP	12.33	30.09	17.49	11.92	32.63	**17.46**	14.06	19.07	16.19
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	12.18	22.26	15.74	11.92	24.14	**15.96**	14.06	14.11	14.09

Table 48: Naive Bayes results.

The most striking observation concerning the Naive Bayes classifier trained on non-oversampled data is that it fails to find even a single VP antecedent. F-measure for NP antecedents, on the other hand, is reasonably well at least for the 2-Trained and 3-Trained classifier, mainly due to a recall in the high 20s to mid 30s in combination with a precision around 10. The best F-measure for ALL is 16.60, produced by the 2-Trained classifier in the setting -tipster, +it-filter, -filter. This setting also produces the best F-measure for NP (18.90). Generally, the best performance for ALL is coupled with the

				2-Trained			3-Trained			4-Trained		
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	9.20	32.20	14.31	9.81	40.68	15.80	12.33	27.54	17.04
			VP	2.22	1.32	1.65	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	8.84	24.14	12.94	9.81	30.09	14.79	12.33	20.38	15.37
		+tipster	NP	9.28	32.63	14.45	10.20	41.10	16.34	11.35	26.70	15.93
			VP	2.33	1.32	1.68	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	8.94	24.45	13.09	10.20	30.41	15.28	11.35	19.75	14.42
	+it-filter	-tipster	NP	10.07	32.63	15.39	11.11	39.83	17.38	12.83	25.00	16.95
			VP	2.33	1.32	1.68	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	9.65	24.45	13.84	11.11	29.47	16.14	12.83	18.50	15.15
		+tipster	NP	9.99	32.20	15.25	10.87	38.14	16.92	12.60	25.85	16.94
			VP	2.44	1.32	**1.71**	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	9.60	24.14	13.74	10.87	28.21	15.69	12.60	19.12	15.19
+filter	-it-filter	-tipster	NP	10.73	38.56	16.79	10.50	41.95	16.79	12.40	26.27	16.85
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.52	28.53	15.37	10.50	31.03	15.69	12.40	19.44	15.14
		+tipster	NP	11.10	38.56	17.24	11.33	45.76	**18.17**	13.02	27.97	17.77
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.92	28.53	15.80	11.33	33.86	**16.98**	13.02	20.69	15.98
	+it-filter	-tipster	NP	12.62	40.68	**19.26**	11.57	40.68	18.01	13.15	23.73	16.92
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	12.36	30.09	**17.52**	11.57	30.09	16.71	13.15	17.56	15.03
		+tipster	NP	11.92	37.29	18.07	11.49	41.10	17.96	14.32	26.27	**18.54**
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	11.73	27.59	16.46	11.49	30.41	16.68	14.32	19.44	**16.49**

Table 49: Naive Bayes results, 3-fold oversampling.

best NP performance, which is not surprising since no VP antecedents are found by the classifier.

The most obvious effect of 3-fold oversampling is that now the Naive Bayes classifier finds VP antecedents at least with the 2-Trained classifier in the setting -filter. Thus, oversampling has a positive effect on F-measure for VP in four out of 24 settings. For ALL, a positive effect can be observed in 20 settings, which just happen to be the ones in which there is no effect for VP antecedents. The same correlation can be observed for NP antecedents: They also improve through 3-fold oversampling, except in settings where there is improvement for VP antecedents. The best F-measure for ALL is 17.52, yielded by the 2-Trained classifier in the setting -tipster, +it-filter, and +filter.

6-fold oversampling yields an increase in F-measure over 3-fold oversampling for ALL and NP in only two out of 24 settings, respectively. VP antecedents are found in only two settings (-tipster, -it-filter, -filter and -tipster, +it-filter, -filter) out of 24, and both with a decrease in F-measure compared to 3-fold oversampling. The best F-measure

				2-Trained			3-Trained			4-Trained		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	8.43	34.32	13.53	10.09	44.49	16.45	10.80	27.97	15.58
			VP	1.82	1.32	1.53	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	8.07	25.71	12.29	10.09	32.92	15.44	10.80	20.69	14.19
		+tipster	NP	8.67	34.32	13.85	9.43	42.37	15.43	8.78	22.88	12.69
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	8.26	25.39	12.46	9.43	31.35	14.50	8.78	16.93	11.56
	+it-filter	-tipster	NP	8.80	33.48	13.93	10.76	41.53	17.09	10.88	24.58	15.09
			VP	1.89	1.32	**1.55**	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	8.41	25.08	12.60	10.76	30.72	15.94	10.88	18.18	13.62
		+tipster	NP	9.54	34.75	14.96	10.74	41.95	**17.10**	10.69	24.15	14.82
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	9.06	25.71	13.40	10.74	31.03	**15.96**	10.69	17.87	13.38
+filter	-it-filter	-tipster	NP	9.25	37.29	14.83	9.08	39.83	14.79	11.17	25.85	15.60
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	9.04	27.59	13.62	9.08	29.47	13.89	11.17	19.12	14.10
		+tipster	NP	10.07	40.25	16.12	9.78	42.80	15.92	11.71	27.54	**16.44**
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	9.93	29.78	14.89	9.78	31.66	14.94	11.71	20.38	**14.87**
	+it-filter	-tipster	NP	10.88	39.41	17.05	10.55	40.68	16.75	11.89	23.73	15.84
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.65	29.15	15.60	10.55	30.09	15.62	11.89	17.56	14.18
		+tipster	NP	10.98	39.41	**17.18**	10.53	40.68	16.73	11.90	24.15	15.94
			VP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			ALL	10.83	29.15	**15.79**	10.53	30.09	15.60	11.90	17.87	14.29

Table 50: Naive Bayes results, 6-fold oversampling.

with 6-fold oversampling for ALL is only 15.96, yielded by the 3-Trained classifier with the setting +tipster, +it-filter, -filter.

Thus, the best result of the Naive Bayes classifier for ALL is precision 12.36, recall 30.09, and F-measure 17.52. It is produced by the 2-Trained classifier (3-fold oversampling) in the setting -tipster, +it-filter, +filter. This setting, as well as the majority of settings for the Naive Bayes classifier, fails to find even a single VP antecedent.

Logistic Regression

				2-Trained			**3-Trained**			**4-Trained**		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	16.67	19.49	**17.97**	15.93	18.22	17.00	13.85	17.37	**15.41**
			VP	6.52	7.90	7.14	8.77	6.58	7.52	2.27	1.32	1.67
			ALL	14.13	16.30	**15.14**	14.68	15.05	14.86	12.35	13.17	**12.75**
		+tipster	NP	15.36	18.22	16.67	18.12	22.03	19.89	13.78	16.53	15.03
			VP	7.22	9.21	8.09	7.94	6.58	7.19	2.22	1.32	1.65
			ALL	13.26	15.67	14.37	16.29	17.87	17.04	12.20	12.54	12.37
	+it-filter	-tipster	NP	17.27	18.22	17.73	16.74	17.37	17.05	12.45	13.98	13.17
			VP	6.59	7.90	7.19	8.48	6.58	7.41	2.22	1.32	1.65
			ALL	14.41	15.36	14.87	15.13	14.42	14.77	10.97	10.66	10.81
		+tipster	NP	16.40	17.37	16.87	20.23	22.88	**21.47**	13.90	15.25	14.55
			VP	7.07	9.21	8.00	7.58	6.58	7.04	2.08	1.32	1.61
			ALL	13.75	15.05	14.37	17.72	18.50	18.10	12.05	11.60	11.82
+filter	-it-filter	-tipster	NP	16.09	17.80	16.90	18.53	20.34	19.39	12.44	11.44	11.92
			VP	6.52	7.90	7.14	13.79	10.53	11.94	5.26	2.63	3.51
			ALL	13.60	15.05	14.29	17.67	17.56	17.61	11.37	9.09	10.11
		+tipster	NP	15.67	17.80	16.67	19.33	22.03	20.59	13.22	12.71	12.96
			VP	7.48	10.53	**8.74**	13.43	11.84	**12.59**	4.62	3.95	**4.26**
			ALL	13.33	15.67	14.41	18.16	19.12	**18.63**	11.30	10.35	10.80
	+it-filter	-tipster	NP	15.83	16.10	15.97	17.87	17.80	17.83	11.52	9.32	10.30
			VP	6.45	7.90	7.10	13.12	10.53	11.68	5.26	2.63	3.51
			ALL	13.21	13.79	13.50	16.89	15.67	16.26	10.48	7.52	8.76
		+tipster	NP	16.10	16.10	16.10	20.82	21.61	21.21	12.38	10.59	11.42
			VP	7.41	10.53	8.70	11.27	10.53	10.88	4.55	3.95	4.23
			ALL	13.37	14.42	13.88	18.67	18.50	18.58	10.45	8.78	9.54

Table 51: Logistic Regression results.

The Logistic Regression classifier in Table 51 is the first to be able to find VP antecedents in all 24 settings. The best F-measure for ALL is 18.63, yielded by the 3-Trained classifier in the setting +tipster, -it-filter, +filter. It is striking that this setting produces the best F-measures for VP for all three classifiers (2-, 3-, and 4-Trained).

As can be seen in Table 52, 3-fold oversampling causes the expected increase in recall for

				2-Trained			**3-Trained**			**4-Trained**		
Setting			**Ante**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
-filter	-it-filter	-tipster	NP	11.20	29.24	16.20	12.93	28.81	17.85	13.71	20.34	**16.38**
			VP	4.52	9.21	6.06	8.14	9.21	8.64	4.65	2.63	3.36
			ALL	9.86	23.82	13.95	12.26	23.51	16.11	12.72	15.67	**14.05**
		+tipster	NP	10.52	26.70	15.09	14.54	31.78	19.95	13.50	18.64	15.66
			VP	5.95	13.16	8.20	8.60	10.53	**9.47**	1.82	1.32	1.53
			ALL	9.52	22.88	13.44	13.63	26.02	17.89	11.81	14.11	12.86
	+it-filter	-tipster	NP	11.45	27.12	16.10	13.76	28.39	18.53	13.65	18.22	15.61
			VP	4.43	9.21	5.98	7.45	9.21	8.24	4.55	2.63	3.33
			ALL	9.90	22.26	13.71	12.74	23.20	16.44	12.54	14.11	13.27
		+tipster	NP	10.36	24.15	14.50	16.24	32.20	**21.59**	13.85	17.37	15.41
			VP	5.88	13.16	8.13	6.54	9.21	7.65	1.75	1.32	1.50
			ALL	9.31	21.00	12.90	14.44	26.02	**18.57**	11.90	13.17	12.50
+filter	-it-filter	-tipster	NP	12.98	30.09	**18.14**	14.96	30.93	20.17	11.79	14.83	13.13
			VP	5.23	11.84	7.26	3.77	5.26	4.40	4.17	2.63	3.23
			ALL	11.13	25.08	**15.41**	12.96	24.14	16.87	10.73	11.60	11.15
		+tipster	NP	12.18	27.97	16.97	14.78	30.51	19.92	12.24	14.83	13.41
			VP	7.37	18.42	**10.53**	7.14	10.53	8.51	4.17	3.95	**4.05**
			ALL	10.93	25.08	15.22	13.36	25.08	17.43	10.62	11.91	11.23
	+it-filter	-tipster	NP	12.85	27.12	17.44	15.77	29.66	20.59	10.86	12.29	11.53
			VP	4.65	10.53	6.45	3.33	5.26	4.08	4.26	2.63	3.25
			ALL	10.75	22.57	14.56	13.12	23.20	16.76	9.87	9.72	9.80
		+tipster	NP	12.68	26.70	17.19	14.93	27.97	19.47	12.25	13.14	12.68
			VP	6.45	15.79	9.16	7.56	11.84	9.23	3.95	3.95	3.95
			ALL	10.98	23.51	14.97	13.37	23.51	17.05	10.33	10.66	10.49

Table 52: Logistic Regression results, 3-fold oversampling.

				2-Trained			3-Trained			4-Trained		
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	9.88	36.02	15.51	13.66	39.41	20.28	12.99	21.19	16.10
			VP	4.65	10.53	6.45	5.36	7.90	6.38	4.35	2.63	3.28
			ALL	9.01	29.15	13.77	12.48	31.03	17.81	12.07	16.30	**13.87**
		+tipster	NP	9.28	32.63	14.45	12.71	36.02	18.79	13.96	20.76	**16.70**
			VP	5.00	11.84	7.03	5.13	7.90	6.22	1.70	1.32	1.48
			ALL	8.52	26.96	12.94	11.58	28.53	16.47	12.20	15.67	13.72
	+it-filter	-tipster	NP	10.09	32.63	15.41	14.78	37.71	**21.24**	13.43	19.92	16.04
			VP	4.55	10.53	6.35	4.84	7.90	6.00	4.26	2.63	3.25
			ALL	9.05	26.65	13.51	13.09	29.78	**18.18**	12.34	15.36	13.69
		+tipster	NP	9.92	31.36	15.07	14.14	34.75	20.10	14.42	19.49	16.58
			VP	5.50	13.16	**7.75**	4.55	7.90	5.77	1.64	1.32	1.46
			ALL	9.05	26.33	13.47	12.36	27.59	17.07	12.37	14.73	13.45
+filter	-it-filter	-tipster	NP	10.42	33.90	15.94	13.26	35.17	19.26	11.91	16.10	13.69
			VP	4.04	10.53	5.84	5.26	9.21	**6.70**	4.00	2.63	3.18
			ALL	9.11	27.59	13.70	11.86	28.21	16.70	10.84	12.54	11.63
		+tipster	NP	9.90	30.51	14.95	13.49	36.02	19.63	12.06	14.41	13.13
			VP	4.55	13.16	6.76	2.24	3.95	2.86	3.85	3.95	**3.90**
			ALL	8.66	25.71	12.95	11.52	27.59	16.25	10.28	11.60	10.90
	+it-filter	-tipster	NP	11.01	31.78	**16.36**	13.22	31.78	18.68	11.70	13.98	12.74
			VP	4.02	10.53	5.82	4.80	9.21	6.31	4.00	2.63	3.18
			ALL	9.43	26.02	**13.85**	11.50	25.71	15.89	10.54	10.97	10.75
		+tipster	NP	10.41	28.81	15.30	12.65	30.51	17.89	11.86	12.71	12.27
			VP	4.63	13.16	6.85	2.04	3.95	2.69	3.75	3.95	3.85
			ALL	8.98	24.45	13.13	10.48	23.51	14.49	9.91	10.35	10.12

Table 53: Logistic Regression results, 6-fold oversampling.

ALL, NP and VP in almost all settings. However, the associated drop in precision almost exclusively outweighs it, so that the best result yielded without any oversampling is not improved upon. Rather, the best result for the 3-fold oversampled data for ALL is only precision 14.44, recall 26.02, and F-measure 18.57. It is produced by the 3-Trained classifier in the setting +tipster, +it-filter, -filter.

The results for 6-fold oversampling in Table 53 more or less repeat the tendency already observed for 3-fold oversampling: A general increase in recall, but a greater and similarly general decrease in precision fail to bring about a better result. The best result for ALL is precision 13.09, recall 29.78, and F-measure 18.18, produced by the 3-Trained classifier in the setting -tipster, +it-filter, -filter.

The best result that can be reported for the Logistic Regression classifier is precision 18.16, recall 19.12, and F-measure 18.63. It is produced by the 3-Trained classifier (no oversampling) in the setting +tipster, -it-filter, +filter. This setting is also the one producing the best F-measure for VP antecedents.

Instance-Based Learning

				2-Trained			3-Trained			4-Trained		
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	7.37	28.39	**11.70**	9.19	29.24	13.98	8.01	13.14	**9.95**
			VP	4.62	11.84	6.64	4.29	7.90	5.56	5.41	5.26	5.33
			ALL	6.88	23.82	**10.67**	8.42	23.51	12.40	7.79	11.29	**9.22**
		+tipster	NP	6.98	26.70	11.06	8.58	27.54	13.08	7.47	11.86	9.17
			VP	5.13	13.16	7.38	3.62	6.58	4.67	4.94	5.26	5.10
			ALL	6.64	22.88	10.30	7.81	21.94	11.52	7.22	10.35	8.51
	+it-filter	-tipster	NP	6.67	21.61	10.19	8.28	22.46	12.10	7.16	10.17	8.41
			VP	4.81	11.84	6.84	5.67	10.53	**7.37**	5.48	5.26	**5.37**
			ALL	6.30	18.81	9.43	7.81	19.12	11.09	7.09	9.09	7.97
		+tipster	NP	6.18	19.92	9.43	8.06	22.03	11.81	6.75	9.32	7.83
			VP	5.79	14.47	**8.27**	5.07	9.21	6.54	3.75	3.95	3.85
			ALL	6.09	18.18	9.13	7.54	18.50	10.71	6.39	8.15	7.16
+filter	-it-filter	-tipster	NP	7.39	25.85	11.49	10.25	29.24	**15.18**	8.80	11.44	**9.95**
			VP	3.88	10.52	5.67	3.97	7.90	5.29	4.00	3.95	3.97
			ALL	6.68	21.63	10.21	9.10	23.51	13.12	8.09	9.72	8.83
		+tipster	NP	7.15	25.00	11.12	10.13	29.24	15.05	8.53	10.59	9.45
			VP	4.95	13.16	7.19	3.40	6.58	4.48	3.70	3.95	3.82
			ALL	6.71	21.63	10.25	8.94	23.20	12.90	7.73	9.09	8.36
	+it-filter	-tipster	NP	6.65	19.92	9.97	8.96	21.61	12.67	7.31	8.05	7.66
			VP	4.10	10.53	5.90	5.20	10.53	6.96	5.41	5.26	5.33
			ALL	6.09	17.24	9.00	8.16	18.50	11.32	7.16	7.52	7.34
		+tipster	NP	6.68	19.92	10.00	9.34	22.88	13.27	7.66	8.05	7.85
			VP	4.64	11.84	6.67	4.76	9.21	6.28	5.00	5.26	5.13
			ALL	6.23	17.56	9.20	8.41	19.12	11.69	7.30	7.52	7.41

Table 54: IB2 results.

Instance-based learning is also able to find VP antecedents in all 24 settings. The best F-measure for IB2 (Table 54) is 10.67, produced by the 2-Trained classifier in the setting -tipster, -it-filter, -filter.

3-fold oversampling (Table 55) has a positive effect on F-measure for ALL in eight out of 24 settings, for NP in five and for VP in 15. It also produces a best overall F-measure for IB2 of 12.73.

Additional (i.e. 6-fold) oversampling fails to produce a result that is different from the result obtained by 3-fold oversampling.

The best result of the IB2 classifier is precision 9.19, recall 20.69 and F-measure 12.73, produced by the 3-Trained classifier (3-fold oversampling) in the setting -tipster, +it-filter, +filter.

				2-Trained			3-Trained			4-Trained		
Setting			Ante	P	R	F	P	R	F	P	R	F
-filter	-it-filter	-tipster	NP	6.47	25.42	10.32	7.33	23.31	11.16	7.21	12.29	9.09
			VP	4.32	10.53	6.13	4.26	7.90	5.53	4.23	3.95	4.08
			ALL	6.11	21.32	9.50	6.94	19.44	10.23	6.95	10.35	8.31
		+tipster	NP	6.14	24.15	9.79	8.05	25.85	12.27	7.75	12.71	**9.63**
			VP	4.35	10.53	6.15	4.32	7.90	5.58	5.33	5.26	5.30
			ALL	5.84	20.38	9.07	7.56	21.32	11.17	7.54	10.97	**8.94**
	+it-filter	-tipster	NP	6.19	20.34	9.50	7.27	19.49	10.59	6.63	9.75	7.89
			VP	4.57	10.53	6.38	5.00	9.21	6.48	5.63	5.26	**5.44**
			ALL	5.89	17.56	8.82	7.10	17.24	10.06	6.67	8.78	7.58
		+tipster	NP	5.77	19.07	8.86	7.34	19.92	10.73	6.87	9.75	8.06
			VP	4.49	10.53	6.30	5.04	9.21	6.51	5.33	5.26	5.30
			ALL	5.53	16.61	8.29	7.04	17.24	10.00	6.80	8.78	7.66
+filter	-it-filter	-tipster	NP	7.39	26.27	**11.54**	9.45	27.12	**14.02**	7.58	10.59	8.83
			VP	4.12	10.53	5.93	4.61	9.21	6.14	5.48	5.26	5.37
			ALL	6.77	21.94	**10.35**	8.79	22.88	12.70	7.41	9.40	8.29
		+tipster	NP	6.83	24.15	10.64	8.66	25.00	12.87	7.64	10.17	8.73
			VP	4.23	10.53	6.04	4.83	9.21	6.34	5.48	5.26	5.37
			ALL	6.34	20.38	9.67	8.09	21.00	11.68	7.71	9.40	8.48
	+it-filter	-tipster	NP	7.06	21.19	10.59	9.88	23.73	13.95	6.43	7.63	6.98
			VP	4.35	10.53	6.15	5.37	10.53	7.11	5.56	5.26	5.41
			ALL	6.50	18.18	9.57	9.19	20.69	**12.73**	6.50	7.21	6.84
		+tipster	NP	6.36	19.07	9.53	8.92	21.61	12.62	7.12	8.05	7.56
			VP	5.03	11.84	**7.06**	5.52	10.53	**7.24**	5.48	5.26	5.37
			ALL	6.08	16.93	8.95	8.48	19.12	11.75	7.31	7.84	7.56

Table 55: IB2 results, 3-fold oversampling.

7.4.1 Discussion

In this chapter, we discuss the comparative performance of the five systems (one baseline algorithm and four classifiers). We are particularly interested in the statistical significance of observed differences in performance (in terms of precision, recall, and F-measure). There are two types of differences that are potentially relevant: The difference between the performance of each of the four non-baseline systems and the best baseline system, and the difference between the performance yielded by a non-baseline system *with* and *without* a particular feature. In our case, the latter type of difference evaluates whether the contribution of each of the binary features +/-tipster, +/-it-filter, and +/-filter is statistically significant.

We compute significance by means of a one-tailed paired t-test (Hays, 1994). A one-tailed paired t-test can be used to determine whether the differences between a sequence of n pairs of related observations $\{< x_1, y_1 >, < x_2, y_2 >, ..., < x_n, y_n >\}$ are statistically significant, given a particular confidence threshold. The null hypothesis for the t-test is that the observations $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$ come from the same underlying distribution and are thus not significantly different. The t-test produces a t-value which takes into account the means and variances within the sets of individual observations (i.e. $\{x_1, x_2, ..., x_n\}$ and $\{y_1, y_2, ..., y_n\}$) as well as those of the pairwise differences. The higher this t-value is for a given sequence of related observations, the more significant are the differences. More specifically, we use a *paired* t-test because – for each hypothesis – our data consists of five pairs of observations, one pair for each dialog in our data set. When testing whether the non-baseline systems are significantly better than the baseline algorithm, the first observation in each pair is the result produced by the baseline, and the second observation is the result produced by the respective system. On the other hand, when testing the significance of the contribution of a particular binary feature, the first observation is the result *without* this feature and the second observation is the result *with* this feature. In all cases, we use a *one-tailed* t-test because our hypothesis is that the respective second observations in each pair are not just significantly different, but significantly *better*.

An important parameter of the t-test is the number of the degrees of freedom (*df*), which is $n\ (=\ number\ of\ observations) - 1$. In our case, $n = 5$, so that $df = 4$. According to Hays (1994), for a one-tailed t-test with $df = 4$, the minimum t-value for a confidence level of $p <= 0.05$ is 2.132, while for a confidence level of $p <= 0.01$ it is 3.747. For a confidence level of $p <= 0.005$ and $p <= 0.001$ the minimum t-value is 4.604 and 7.173, respectively.

The best results for each classifier are tabulated in Table 56. The statistical significance of the performance differences between baseline and non-baseline systems is marked in the three righthand columns headed P, R and F. A single asterisk (*) means that the marked result is significantly better than the corresponding baseline for a confidence level of $p <= 0.05$, while a double asterisk (**) means that the marked result is significantly better for a confidence level of $p <= 0.01$. Three and four asterisks are used accordingly for $p <= 0.005$ and $p <= 0.001$. The significance of the contribution of the three binary features +/-tipster, +/-it-filter, and +/-filter is encoded in the respective table cells. Each cell contains the letters P, R and F, with either a plus or a minus sign. This sign specifies the direction of the effect, i.e. whether the feature has an overall positive or negative effect on the respective measure. For the baseline algorithm, e.g. the table states that using the it-filter has a positive effect on precision and F-measure for NP, and a negative effect on recall for NP. The effect on precision is statistically significant for $p <= 0.05$, while the effects on recall and F-measure are not significant. For better readability, grey highlighting is used in the table for those cells that contain at least one significant difference.

In Table 56, we see that all classifiers have a significantly better precision than the baseline algorithm in the most important category ALL. For all classifiers except J48 this also results in a significant improvement in terms of F-measure for this category. The only significant improvement in recall over the baseline can be observed for the Naive Bayes classifier. However, this improvement comes at the price of a zero F-measure for VP antecedents, because even in its best result, the Naive Bayes classifier fails to find even a single antecedent of this type. The tipster feature does have significant effect on two classifiers: For J48, it significantly *decreases* both precision and F-measure for NP and ALL, while for the Naive Bayes classifier, it also has a significant negative effect on recall for NP and ALL. Significant effects of the it-filter feature are all positive for the Baseline (significant improvement in precision for NP) and for Naive Bayes (significant improvement in precision and F-measure for NP and ALL). The Naive Bayes classifier is also the only one on which the filter feature does have any significant (and consistently positive) effect.

Thus, the best classifier in terms of F-measure for ALL is Logistic Regression. It has the highest precision and F-measure for both NP, VP, and ALL. What is striking about this classifier is that it is not at all significantly affected by any of the three binary features.

Classifier	Over-sampling	Trained	Search depth	tipster	it-filter	filter	Ante type	P	R	F
Baseline	n/a	n/a	6/3	n/a	+P*/-R /+F	n/a	NP	5.38	25.85	8.91
					-P /n-a/-F		VP	2.04	2.63	2.30
					+P /-R /+F		ALL	5.12	19.75	8.13
J48	6-fold	Core 3	9/7	-P*/-R /-F*	+P /-R /+F	-P /-R /-F	NP	10.12*	25.42	14.48
				+P /-R /-F	-P /-R /-F	+P /+R /+F	VP	2.02	2.63	2.29
				-P*/-R /-F*	+P /-R /+F	-P /-R /-F	ALL	9.22*	20.06	12.64
N. B.	3-fold	Core 2	13/7	-P /-R*/-F	+P*/+R /+F*	+P*/+R*/+F*	NP	12.62**	40.68*	19.26***
							VP	0.00	0.00	0.00
				-P /-R*/-F	+P*/+R /+F*	+P*/+R*/+F*	ALL	12.36***	30.09*	17.52***
Log. Reg.	none	Core 3	9/7	+P /+R /+F	+P /-R /+F	+P /-R /+F	NP	19.33****	22.03	20.59***
				+P /+R /+F	-P /-R /-F	+P /+R /+F	VP	13.43	11.84	12.59
				+P /+R /+F	+P /-R /-F	+P /+R /+F	ALL	18.16***	19.12	18.63**
IB2	3-fold	Core 3	9/7	-P /-R /-F	+P /-R /+F	+P /+R /+F	NP	9.88*	23.73	13.95*
				+P /-R /-F	+P /n-a/+F	+P /+R /+F	VP	5.37	10.53	7.11
				-P /-R /-F	+P /+R /+F	+P /+R /+F	ALL	9.19*	20.69	12.73**

Table 56: Final results.

Although the effect of the tipster feature is consistently positive, it fails to be significant. The effect of the it-filter is even negative, although not significant either.

Apart from yielding the best overall performance of all classifiers employed, the Logistic Regression classifier has a couple of other advantages. First, it does find at least some VP antecedents[56], even if the performance fails to be statistically significant over the Baseline. Second, it yields its best performance when trained on the non-oversampled data. This is an attractive characteristic because it means that this classifier does not require oversampling, including the determination or estimation of the optimal oversampling rates.

7.4.2 Qualitative Performance Analysis

Apart from a mere quantitative, black-box comparison of the performance of the different classifiers in terms of precision, recall, and F-measure, a more qualitatively oriented comparison of their individual resolution decisions is also useful. Among other things, it can provide information about whether some classifiers handle correctly different types of anaphors. If this turned out to be true, the *combination* of classifiers could be a way to improve overall system performance. The qualitative analysis that follows does only have an exemplary character: Full analysis would have to include the detailed analysis of correct as well as incorrect and missing resolutions (*error analysis*). This is beyond the scope of this thesis (but see Chapter 8). In this thesis, we can only provide a brief inspection which does, however, provide some useful insights.

In order to limit the number of cases to be inspected in the following, we only consider the anaphors contained in the key, i.e. in the correct anaphoric chains. We only analyze whether a given anaphor in the key was or was not resolved correctly by a given classifier. Thus, we do not distinguish between recall errors caused by *unresolved* anaphors and those caused by *wrongly resolved* anaphors. This also means that precision errors are left out of the discussion.

As was described in Chapter 7.1, we consider as potentially resolvable pronouns in the test data, i.e. in the key, only those instances of *it*, *this*, and *that* which appear in chains which have *non-pronominal* antecedents. The total number of potentially resolvable pronouns in our test data set (i.e. in core data set 3) is 319. This number is calculated on the

[56] Precision, recall and F-measure for VP translate to 67 of 76 detected VP antecedents, 9 of which are correct.

basis of the anaphoric chain statistics in core data set 3 in Table 26 on page 99, repeated below as Table 57.

		2	3	4	5	6	7	8	9	10	11	12	total
	NP	17	3	2	-	1	-	-	-	-	-	-	23
	PRO	14	-	2	-	-	-	-	-	-	-	-	16
Bed017	VP	6	1	-	-	-	-	-	-	-	-	-	7
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	37	4	4	-	1	-	-	-	-	-	-	46
	NP	14	4	1	1	1	1	-	-	-	1	-	23
	PRO	19	9	2	2	1	-	1	-	-	-	-	34
Bmr001	VP	9	5	-	-	-	-	-	-	-	-	-	14
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	42	18	3	3	2	1	1	-	-	1	-	71
	NP	18	3	3	1	-	-	-	-	-	-	-	25
	PRO	18	1	1	-	-	-	-	-	-	-	-	20
Bns003	VP	14	4	-	-	-	-	-	-	-	-	-	18
	OTHER	-	-	-	-	-	-	-	-	-	-	-	-
	all	50	8	4	1	-	-	-	-	-	-	-	63
	NP	38	5	3	1	-	-	-	-	-	-	-	47
	PRO	21	4	-	1	-	-	-	-	-	-	-	26
Bro004	VP	8	1	1	-	-	-	-	-	-	-	-	10
	OTHER	2	1	-	-	-	-	-	-	-	-	-	3
	all	69	11	4	2	-	-	-	-	-	-	-	86
	NP	37	7	1	-	-	-	-	-	-	-	-	45
	PRO	15	3	1	-	-	-	-	-	-	-	-	19
Bro005	VP	8	1	-	1	-	-	-	-	-	-	-	10
	OTHER	3	-	-	-	-	-	-	-	-	-	-	3
	all	63	11	2	1	-	-	-	-	-	-	-	77
	NP	124 76.07	22	10	3	2	1	-	-	-	1	-	163 47.52
	PRO	87 75.65	17	6	3	1	-	1	-	-	-	-	115 33.53
Σ	VP	45 76.27	12	1	1	-	-	-	-	-	-	-	59 17.20
	OTHER	5 83.33	1	-	-	-	-	-	-	-	-	-	6 1.75
	all	261 76.09	52	17	7	3	1	1	-	-	1	-	343 100.00

Table 57: Anaphoric chain statistics in core data set for $n = 3$.

The number of resolvable anaphors is calculated by multiplying the frequency of chains of a particular length with this length minus 1, and summing over all products. For chains with NP antecedents, e.g. the calculation is thus as follows:

$$124 * 1 + 22 * 2 + 10 * 3 + 3 * 4 + 2 * 5 + 1 * 6 + 1 * 10 = 236$$

For VP and Other antecedents, the above calculation yields 76 and 7 resolvable anaphors, respectively, resulting in a total of 319.

We first determine those anaphors which were correctly resolved by a given number of classifiers (0 to 5). In doing so, we distinguish between the type of the correct antecedent, i.e. NP, VP, or Other. The results can be found in Table 58.

Correctly	NP			VP			OTHER			Total	
resolved by	#	%	% res.	#	%	% res.	#	%	% res.	#	% Total
0	100	42.37	59.52	62	81.58	36.90	6	85.71	3.57	168	52.66
1	59	25.00	85.51	10	13.16	14.49	-	-	-	69	21.63
2	25	10.59	83.33	4	5.26	13.33	1	14.29	3.33	30	9.40
3	33	13.98	100.00	-	-	-	-	-	-	33	10.35
4	11	4.66	100.00	-	-	-	-	-	-	11	3.45
5	8	3.39	100.00	-	-	-	-	-	-	8	2.51
Σ	236	73.98		76	23.82		7	2.19		319	100.00

Table 58: Numbers of anaphors resolved correctly by 0 to 5 classifiers.

This table is to be read as follows: The line **Correctly resolved by 0** contains the number of anaphors that none of the classifiers resolved correctly. This number is 168, which constitutes 52.66% of all resolvable anaphors (rightmost column). Of these 168 anaphors, 100 anaphors have NP antecedents (# column under NP), and they constitute 42.37% of all anaphors with NP antecedents (% column under NP), and 59.52% of all anaphors resolved correctly by 0 classifiers (% res. column under NP). Note that no distinction is made between those cases where a classifier left an anaphor *unresolved* and those cases where it resolved it to an *incorrect* antecedent: All that is said is that the stated number of anaphors does not appear in any classifier's list of correctly resolved anaphors. The other lines (**1** - **5**) are to be interpreted accordingly.

7.4.2.1 Analysis of Anaphors of Class *Correctly resolved by 0*

It is not surprising in view of the quantitative evaluation results reported in Table 56 on page 211 above that the anaphors which no classifier is able to correctly resolve constitute the majority of cases (52.66%). This class also includes the vast majority of anaphors with VP antecedents: Of a total of 76 anaphors with VP antecedents in the key, as many as 62 (81.58%) are not correctly resolved by even a single classifier.

In order to get an idea of the nature of the anaphors that were not correctly resolved by any of the classifiers, we randomly selected five anaphors with NP and five with VP antecedents from this set. In doing so, we restricted ourselves to anaphors in chain position 2, i.e. those anaphoric instances of *it*, *this*, and *that* that *directly follow* the non-pronominal antecedent in the anaphoric chain. This way, we can analyze the classifiers' ability to pick a non-pronominal antecedent. Note that this does not seriously restrict the set of examples to choose from, because 76.09% of all anaphoric chains in the test data set consist of one antecedent and only one anaphor, anyway.

In the following examples, the correct antecedent for each anaphor (as indicated by at least three of our four annotators) is shown by coindexing. Some examples (mostly those involving NP antecedents) are also briefly discussed. Many examples, however, remain undiscussed because no clues are available as to why they were handled correctly or incorrectly by the respective classifier(s).

Anaphors with NP antecedents not correctly resolved by any classifier:

Example 1

MN059: [...] As a matter of fact, the only thing that m- apparently really works out so far are [library ordering codes]$_i$, which are very, very coarse grain, so you have some like,

MN059: science, biology, and then - But [that]$_i$'s really all that we have at the moment. [...] (Bed017)

This is an example of erroneous preprocessing: Inspection of the underlying data showed that *ordering* was incorrectly POS-tagged as a present particple (VBG) during parsing (Chapter 6.1.5), preventing the noun phrase *library ordering codes* to be detected as a potential antecedent during chunking (Chapter 6.1.6).

Example 2

ME025: Well no, O_K, the question I still have about what we 're going to do is
ME013: O_K.
ME013: Right.
ME025: what is - what - eh - O_K , so we can get the equipment set up so that
ME025: basically we're in a situation where we can - *Pause* with, you know, ten minutes of preparation, we can bring *Pause* nine people in here and do [a multi- channel recording]$_i$.
ME025: Um.
ME025: What are - wh- When are we - Are we going to do [that]$_i$? [...] (Bmr001)

Here, the error is in the preprocessing as well. The disfluency detection (Chapter 6.1.4) removes *multi-* because it considers it as a word fragment. Note, however, that the actual error is in the original data, because it is the incorrect space character after *multi-* which causes the tokenization to fail to extract the correct *multi-channel*.

Example 3

MN082: *inbreath*
MN082: So, my point of view - I will
MN082: doc- I will
MN082: documented everything
MN082: when I will leave here in the sense that
MN082: potential successor as a group leader
MN082: can hook in with these things
MN082: and that
MN082: [these web pages]$_i$ will be *updated* every time when it is necessary. We tried to do it,
MN082: so that [it]$_i$ must be at least updated only when really major changes happen. (Bns003)

There are several possible reasons why this anaphor was not correctly resolved by any classifier: The plural noun phrase *these web pages* disagrees in number with the singular pronoun *it*. This definitely prevents the resolution in the setting where number disagreement is used as a hard constraint (Chapter 7.2.4). Another possible reason could be the presence of two instances of *it* between the anaphor to be resolved (which is also an instance of *it*) and the noun phrase antecedent.

Example 4

ME013: *Pause* How are we doing on [the
ME013: *Pause* resources]$_i$? Disk, and -
MN007: I think we're alright, um, *Pause* not much problems with [that]$_i$.
(Bro004)

Here, inspection of the underlying data showed that the preprocessing can be ruled out as an error source despite the fact that the NP antecedent occurs at a segment break and has an embedded *Pause* element, since the chunker correctly identified the NP *the resources*.

Example 5

FN002: And then w- with [the first configuration]$_i$, I f-
ME013: *laugh*
ME013: Yeah.
FN002: I am found [that]$_i$
FN002: work,
FN002: uh, doesn't work -
FN002: uh, well,
FN002: *work*, but is better, the second configuration. (Bro005)

While in some of the previous examples preprocessing errors caused the antecedent to be missing, this example is unresolvable because the anaphor was not identified as a chunk. This is because the anaphor *that* was incorrectly tagged as a complementizer during preprocessing. This, in turn, is mainly caused by the ill-formedness of the containing utterance by a non-native speaker of English.

Anaphors with VP antecedents not correctly resolved by any classifier:

Example 1

ME010: And, is now
ME010: doing all the politics for CITRIS, but also, [has]$_i$ a uh, a lot of interest in
ME010: uh, actually doing things for society, so digital divide and stuff like that. So [that]$_i$'s
ME010: s- interesting to me but maybe not to you. [...] (Bed017)

Example 2

ME025: Right. But I mean if we - if we want *Pause* Jerry's group to *Pause* use it then we probably @@ won't ask them [to read]$_i$ any numbers, right? Cuz.
ME011: Right. Well, I - I bet they would be willing to do [it]$_i$ the first few times, just for the novelty. (Bmr001)

Example 3

MN082: I would like to say, earliest stage [will]$_i$ be end of January,
MN082: so [that]$_i$ means *Pause* we have to get over six weeks and - (Bns003)

Example 4

ME013: *Pause* Notice how I said somebody and
ME034: *laugh*
ME013: *laugh*
ME013: [turned]$_i$ my head your direction. [That]$_i$'s one thing you don't get in these recordings. [...] (Bro004)

Example 5

ME013: So, i- it's precisely given that model you [can]$_i$ very simply affect, uh, the s- the strength that you apply the features.
ME013: That was - [that]$_i$ was, uh, Hari's suggestion. (Bro005)

7.4.2.2 Analysis of Anaphors of Class *Correctly resolved by 1*

Another interesting fact to be read out of Table 58 above is that 21.63% of all resolvable anaphors are correctly resolved by exactly one classifier only. This indicates some degree of disjointness of the individual classifiers' capabilities, which is further underlined by the fact that only 2.51% of all anaphors are resolved by all five classifiers.

In order to investigate the apparent strengths of the individual classifiers, the 69 anaphors which were correctly resolved by exactly one classifier can be broken down according to the classifier involved. The result can be found in Table 59.

Classifier	NP	VP	Other	Total
Baseline	12	2	-	14
J48	15	2	-	17
N.B.	17	-	-	17
Log. Reg.	7	4	-	11
IB2	8	2	-	10
Σ	59	10	-	69

Table 59: Anaphors resolved correctly by exactly one classifier.

The table shows that the anaphor subsets that are correctly resolved by exactly one classifier are of sizes between 10 and 17. In view of the simplicity of the Baseline, it is surprising that it still accounts for 14 anaphors that are not correctly resolved by any other classifier. Another observation is that the Logistic Regression classifier resolves the most anaphors with VP antecedents correctly. While the rate is low in absolute terms (4), it is considerably more than for the other classifiers, which all handle 2 resp. none (Naive Bayes). However, the number of anaphors with NP antecedents resolved correctly by the Logistic Regression classifier alone is only 7.

We again complement the numerical evaluation with a discussion of some randomly selected examples. This time, we select for each classifier two anaphors with NP antecedents and two with VP antecedents, and again, with one exception, we only consider anaphors at chain position 2.

Anaphors with NP antecedents only correctly resolved by Baseline:

Example 1

FN050: And one is
FN050: basically how desirable [a site]$_i$ is
FN050: meaning, um,
FN050: how good [it]$_i$ matches the needs of a user. [...] (Bed017)

Example 2

ME013: *Pause* Because *anyway* when we go to
ME013: *Pause twice* as much *data*
ME013: *Pause* and have the *same* number of *parameters,*
ME013: *Pause* particularly when it's twice as much [data]$_i$ *and* [it]$_i$'s quite *diverse,*
ME013: *Pause* um, I wonder if having twice as many parameters would help. (Bro004)

Anaphors with VP antecedents only correctly resolved by Baseline:

Example 1

ME013: It sounded like they [were]$_i$ roughly equal?
ME013: Is [that]$_i$ about right? (Bro004)

Example 2

ME006: Just [listening]$_i$. *laugh*
ME013: *laugh* Well I *figured* [*that*]$_i$, but - *laugh* (Bro005)

In examples 1 and 2 above, the Baseline algorithm is able to select the correct VP antecedent because the **ArgumentOf** constraint (cf. Chapter 6.2.2) causes it to skip the only incorrect interferring VP antecedent candidate (viz. *is* and *figured*).

Anaphors with NP antecedents only correctly resolved by J48:

Example 1

ME013: [...] Uh, You get [something]$_i$ from the -
ME013: the other site at one point and you work really hard on making [it]$_i$ better with rescoring. (Bro005)

Example 2

MN021: Well, th- they have uh, a - [a white booklet]$_i$,
MN021: you know, probably -
MN082: Yeah, [that]$_i$ is from nineteen ninety-one or nineteen ninety-two, you know. (Bns003)

Anaphors with VP antecedents only correctly resolved by J48:

Example 1 and 2

MN081: M- [move]$_i$ it on the normal ICSI,
FN083: Yeah.
MN081: *I_C_S_I*.
FN083: Maybe.
MN081: Maybe every group has somewhere hidden, uh *Pause*
MN082: We can *do* [that]$_i$.
MN081: a gro- uh
MN081: deep *Pause*
MN082: Yeah, but the *point* is that nobody feel *responsible* for [that]$_i$. (Bns003)

The two sample anaphors appear in the same chain, which is why here we provide a three-element chain as example.

Anaphors with NP antecedents only correctly resolved by Naive Bayes:

Example 1

ME013: *outbreath* O_K,
MN017: Hmm.
ME013: *outbreath* so if we take uh
ME013: um *Pause*
ME013: let's see
ME013: *Pause* [P_L_P]$_i$
ME013: *Pause* uh with on-line *Pause* normalization and *Pause* delta-del- so [that]$_i$'s this thing you have circled here
ME013: *Pause* in the second column,
MN007: Yeah.
ME013: um
ME013: *Pause* and ”multi-English” refers to what? (Bro004)

Example 2

MN082: So maybe we have to *makes noise* rewrite something, have new ideas going forward more in
MN082: certain kind of really telecommunication network and the routing and Q_S stuff,
MN082: and remove, in principle, [the Multicast stuff]$_i$
MN082: which, in principle, a pity because I think [that]$_i$'s a really really good eh -
MN082: uh *Pause*
MN082: mmm
MN082: source of research work [...] (Bns003)

Anaphors with NP antecedents only correctly resolved by Logistic Regression:

Example 1

ME013: Well, I think this is what you were explaining. [The *noise* condition]$_i$ is the same -
ME013: [It]$_i$'s the same uh Aurora noises
ME006: *mike noise*
MN007: Yeah.
ME013: uh, in all three cases
MN007: Yeah.
ME013: *Pause* for the training. (Bro004)

Example 2

ME025: And *Pause* [the math]$_i$ is too difficult for me to understand but it's - but [it]$_i$'s there. (Bmr001)

Anaphors with VP antecedents only correctly resolved by Logistic Regression:

Example 1

ME025: Jim [is]$_i$ busy at the moment. [That]$_i$'s the *problem*. [...] (Bmr001)

Example 2

MN07: If we exclude English, *Pause* um *Pause* there [is]$_i$ *Pause* not much difference with the *Pause* data with English.
ME013: Aha!
MN07: So.
MN07: Yeah.
ME013: [That]$_i$'s *interesting*. (Bro004)

Anaphors with NP antecedents only correctly resolved by IB2:

Example 1

ME013: [...] *Pause* it seemed like uh it - it might simply be a case that you have [something]$_i$ that is just much more diverse,
ME018: Mm-hmm.
ME013: *Pause* but you have the same number of parameters representing [it]$_i$. (Bro004)

Example 2

MN007: We can try *Pause*
MN007: networks
MN007: with [LogRASTA filtered features]$_i$.
ME013: Maybe.
MN007: Mmm.
ME034: Would you be using on-line normalization with [that]$_i$? (Bro005)

Anaphors with VP antecedents only correctly resolved by IB2:

Example 1

ME006: It sometimes, actually, [depends]$_i$ on what features you're using .
ME013: Yeah.
ME013: But - but i- it sounds like -
ME006: Um, but -
ME013: *Pause* I mean. [That]$_i$'s interesting because [...] (Bro004)

Example 2

MN021: Well, they are - they [are]$_i$ using
MN021: the - the - the same idea of labeling *Pause* packets
MN082: Yeah.
MN021: *inbreath*
MN082: Yeah.
MN021: with a -
MN082: Yeah.
MN021: with a
MN021: content descriptor or something.
MN035: [That]$_i$'s right. (Bns003)

7.5 Resolution Experiments with Idealized Data

The result reported in Table 56 on page 211 is the final result for the realistic task, i.e. not assuming any non-trivial knowledge sources. In this setting, however, there are a couple of error sources outside of our system, viz. in the preprocessing.

One possible source of error is the imperfect detection of non-referential *it*. As described in Chapter 6.1.1, the component that we use has a performance of precision 80.0, recall 60.9, and F-measure 69.2. Due to the rather low recall, it could thus be expected that some errors arise from non-referential instances of *it* that are wrongly added to anaphoric chains. According to our evaluation scheme (Chapter 7.1), this will cause a precision error unless the initial element in the chain is itself a pronoun, in which case the non-referential *it* is treated as unresolved. Along the same lines, the imperfect precision could cause instances of *it* to be filtered out, even though they are actual referential instances. If an instance of *it* that is part of a coreference chain is removed in this way, it will cause a recall error.

Another weak point in our system which could be expected to cause errors is the shallow disfluency detection and removal procedure. As was described in Chapter 6.1.4, speech disfluencies are recognized mainly on the basis of mere repetition detection. Since more complex forms of e.g. self-repairs or abandoned utterances are not detected by this method, they remain in the data, potentially causing spurious links and other types of errors.

In a set of alternative experiments, we wanted to eliminate the effect of these two imperfect preprocessing components in order to see what the performance of our system would be with somewhat idealized, cleaner input data. By doing this, we make the condition under which our system operates slightly more similar to the conditions used by Strube & Müller (2003) and Byron (2004). For the alternative experiments we only used the Logistic Regression classifier, because it was the one with the best overall performance. Note that this classifier yielded its best result in a setting which did *not* use the automatic detection of non-referential *it* (-it-filter).

For the manually improved detection of non-referential *it*, we modified the experimental setup as follows. From the four manual annotations produced by annotators 1 and 2 during the first and by annotators 3 and 4 during the second data collection (Chapter 4.1 and Chapter 4.2, respectively), we automatically created a markable level *manual_nonref_it* which contained all instances of *it* that at least three annotators classified as non-referential.[57] The definition of *non-referential* included all instances of *it* that were annotated as discarded, extrapos, or prop-it. It did *not* include vague pronouns.[58] During testing, the information from the markable level *manual_nonref_it* was used in the same was as that from the level *nonref_it* during normal testing with the feature it-filter activated: By skipping all instances of *it* for which the level *manual_nonref_it* contained a markable.

Likewise, an improved disfluency detection and removal was realized in the following way. From a related experiment (described in Strube et al. (2007)), data was available in which one annotator (female undergrad student, American English native speaker) had marked words and word sequences that belonged to speech disfluencies. This data, which was available in the form of a markable level (*manual_disfluencies*), was utilized as a filter during testing by skipping all expressions (i.e. pronoun anaphors and NP and VP antecedents) which were identical to or embedded into a markable on that level.

The results of our experiments with idealized data are summarized in Table 60. Only the variable parameters are shown in the table. All other settings are identical to the ones that produced the best result for the Logistic Regression classifier in the previous chapter. This result is given in the first row of the table (-it-filter, auto disfluency detection).

The statistical significance of the pairwise differences between the first result (fully automatical, i.e. -it-filter, auto disfl. detect.) and the five other results produced by different

[57] The rationale for requiring agreement of at least *three* annotators is the same as that applied in the context of the creation of the core data sets, cf. Chapter 4.2.4.

[58] As was argued in Chapter 4.1, vague pronouns are indistinguishable from normal, referential pronouns on the basis of mere surface features. Therefore, it seems unrealistic to assume that vague pronouns can be pre-filtered in the same way as discarded or extraposed instances of *it* resp. as prop-it.

[59] The t-value of this F-measure is 2.12, which means that it misses significance at the 0.05 level only very narrowly, the threshold being at 2.132.

it-filter	disfl. detect.	Ante	P	R	F
-	auto	NP	19.33	22.03	20.59
		VP	13.43	11.84	12.59
		ALL	18.16	19.12	18.63
	manual	NP	22.31*	24.58	23.39*
		VP	10.15	9.21	9.66
		ALL	19.76	20.38	20.06
+	auto	NP	20.82	21.61	21.21
		VP	11.27	10.53	10.88
		ALL	18.67	18.50	18.58
	manual	NP	22.41*	22.88	22.64
		VP	11.27	10.53	10.88
		ALL	19.87*	19.44	19.65
manual	auto	NP	21.51*	22.88	22.18
		VP	11.59	10.53	11.03
		ALL	19.38*	19.44	19.41
	manual	NP	22.82*	23.31*	23.06*
		VP	10.00	9.21	9.59
		ALL	19.94*	19.44	19.68[59]

Table 60: Final results (idealized input).

settings is given in the columns headed P, R and F. We again use grey highlighting to mark table cells with significant differences.

7.5.1 Discussion

An immediately obvious result is that idealized input only yields an improvement of 1 to 3 percent. This can either mean that the two identified error sources (filter for non-referential *it* and disfluency detection) are not responsible for a large number of errors, or that even the manually enhanced data still contains too many errors. With respect to the effect of using automatically obtained vs. manually improved data, no clear trend for the absolute values for precision, recall, and F-measure can be observed. When only *statistically significant* effects are considered, however, the following interesting observations can be made.

First, the table nicely shows that manual filtering of non-referential *it* and manual disfluency detection are mainly beneficial for resolution *precision*: In the P column, there are seven instances of significant improvements over the fully automatic setting, while there is only one such instance in the R column, and only two in the F column.

Second, the table also shows that the setting using both manual filtering of non-referential *it* and manual disfluency detection yields significant improvements in four categories: P, R, and F for NP antecedents, and P for ALL. This improvement is more consistent than the improvements in any of the other settings, which are only significant for two categories (mostly P for NP and ALL). Thus, our results show that improving the quality of the input data by means of manual preprocessing does have a small and statistically significant positive effect on pronoun resolution.

The gains yielded by manual preprocessing, however, are exclusively in the NP and ALL categories, while the best results for VP are produced in the setting without any manually enhanced data. From the top to the bottom in Table 60, the absolute performance figures for VP even decrease consistently (if not significantly). While it is unlikely that the improved data quality has a direct detrimental effect on the resolution of discourse-deictic pronouns, a more plausible explanation is that the performance increase in the resolution of pronouns with NP antecedents comes at the expense of the performance of the resolution of discourse-deictic pronouns, i.e. those with VP antecedents.

7.6 Chapter Summary

In this chapter we integrated several lines of work from previous chapters of this thesis into a running system for spoken dialog pronoun resolution. The chapter began with the definition of an evaluation measure. As was made clear from the outset (Chapter 1), pronoun resolution in this thesis is performed in the context of a practical application for which the commonly used evaluation measures are inappropriate. Consequently, we defined a more appropriate evaluation measure which takes into account important properties of anaphoric chains, in particular the nature of the chain-initial element. Next, some experimental parameters had to be defined. This process included, on the one hand, the identification and operationalization of rather standard parameters like e.g. the oversampling rate that was used to counter class imbalance, the treatment of incompatibility in terms of number, gender, and person (constraint vs. feature), the application of a filter for non-referential *it*, and the choice of machine learning classifier. On the other hand, since only little work has so far been done on spoken dialog pronoun resolution as attempted in this thesis, less common issues also needed to be addressed. These issues included e.g. the following:

- A data set had to be selected to serve as training and test data. As was discussed in Chapter 4.2, consensus-based creation of a gold-standard data set was not an option for this thesis, so that three core data sets were created automatically instead. For training, the solution was to turn the choice of training data set into an experimental parameter and perform several runs, one with each of the three core data sets. For testing, in contrast, one and the same data set was used throughout for the sake of comparability of the results.

- A way had to be found to control the creation of NP and VP antecedent candidates. In the absence of well-defined utterance boundaries as a means to limit the antecedent search depth, the solution here was to rely on the *temporal* distance between anaphor and potential antecedent. In addition, a criterion had to be found for when VP antecedents should be created. On the basis of insights gained from the literature (Chapter 2.3) and observations in our own data sets (Chapter 4.2.5.2), the solution was to create VP antecedent candidates for all instances of *that* and for all instances of either *it*, *this*, or *that* which appeared as the grammatical object of a form of the verb *do*.
- A resolution algorithm had to be selected. While some algorithms do exist in the literature that are particularly tailored for pronouns with NP and VP antecedents (Chapter 5.2), none of them can easily be implemented to be executed in a fully automatic fashion. Since automatic applicability is another requirement of this thesis that directly results from its application-oriented motivation, an algorithm had to be found which could be fully implemented while at the same time allowing for the representation of the diverse descriptive features and relations that we considered important (Chapter 6.2). We deliberately chose a simple and well-understood algorithm, binary mention-pair classification.

The actual experiments targeted two settings: The first was a realistic, real-life setting which was characterized by there being no manually prepared data. From the application-oriented motivation of this thesis, the set of experiments in this setting was the more important one, as it would yield a practically usable pronoun resolution system for spoken dialog. The performance of the system using the best classifier (Logistic Regression) in the best parameter setting amounted to a precision of 18.16, a recall of 19.12 and an F-measure of 18.63 for the category ALL, i.e. when both NP and VP antecedents are considered. This result is rather low, but it still constitutes a highly significant improvement over a recency-based baseline system. There are no other systems yet with a comparable degree of implementation and automation with which our system could be contrasted. A qualitative analysis of (some of) the results of all classifiers employed provided several insights. First, erroneous preprocessing was one source of error that caused some anaphors to be not correctly resolved at all. Other error sources include ill-formed input on the technical level (wrong tokenization), but also ungrammatical utterances, e.g. by non-native speakers. It was also found that more than 20% of all correctly resolved anaphors were handled correctly by one classifier only, while the rate of anaphors that all classifiers resolved correctly was only 2.5%. Generally, a considerable amount of anaphors with NP antecedents (42.37%) and the vast majority of those with VP antecedents (81.58%) were not resolved correctly at all.

The second set of experiments was more diagnostic in nature. In these experiments, we used idealized data, i.e. data in which both the detection of non-referential *it* and the detection and removal of disfluencies were performed manually. The most striking result of these experiments was that while there indeed was a significant improvement over the fully automatic experiments if the most idealized data was used, this improvement was only in the range of 1 to 3 percent. This can clearly be taken as evidence that spoken dialog pronoun resolution might crucially hinge on the solution of problems that are far less superficial (see Chapter 8). However, it does not invalidate the usefulness of the fully automatic system.

8 Conclusions and Future Work

This thesis had as its point of departure the following observations:

- The pronouns *it*, *this* (to a lesser degree), and *that* are very frequent in spoken dialog.
- Apart from being normal anaphors with NP antecedents, these pronouns are also often discourse-deictic anaphors, i.e. they have antecedents that are not noun phrases.
- The high frequency of these pronouns, and the fact that by being anaphoric they convey important information, makes them relevant objects of investigation in spoken dialog.

The extrinsic motivation for a computational treatment of *it*, *this*, and *that* in spoken dialog came from a real-world application scenario, viz. extractive summarization. The actual extrinsic evaluation, though, was beyond the scope of this thesis (see Mieskes (2008)). For the purposes of this thesis, evaluation was restricted to intrinsic evaluation. The most important result of this thesis with respect to this evaluation is that the pronouns *it*, *this*, and *that* in spoken multi-party dialog to a large extent defy an automatic resolution with the means employed here. The best result obtained by a machine-learning based classifier is precision 18.16, recall 19.12, and F-measure 18.63. While this result is significantly better than the best result produced by a recency-based Baseline system (precision 5.12, recall 19.75, and F-measure 8.13), it fails to be even remotely in the range that is achieved by even the weakest (or earliest) implemented resolution systems for written language.

When we try to explain this result, several points come to mind. Some of these points can directly be translated into questions for future research, of which this thesis offers quite a few.

It is in the nature of spontaneous spoken language (and in sharp contrast to written text) that pronouns are more often vague or ambiguous. Among other things, this is related to performance (vs. competence) issues resulting from the situation of speech production. The quote from Sinclair (2004, p.13) which served as the motto of this thesis succinctly makes this point: ”People do not remember the spoken language exactly and so they cannot refer back to it in quite the simple way that they can with the written language." In the annotation experiments performed for this thesis, this can be regarded as a cause of low agreement on several levels. Agreement (in terms of % overlap) among annotators was low (approx. 50%) for the task of identifying which pronouns were or were not referential (i.e. were or were not member of a coreference chain). Among the pronouns that *were* identified as referential by all annotators, agreement (in terms of Krippendorff's α) was low (approx. .50) for the task of identifying the correct NP or VP antecedent. This situation, in combination with the fact that consistent data sets for training and testing were required, prompted the employment of a rather restrictive data consolidation approach. While this approach produced consistent data sets, this very strict definition might turn out to be overly strict. For future research, more flexible approaches should be evaluated. One obvious approach would be to accept ambiguity as a fact (rather than trying to eliminate it), by allowing *all* different annotations into the data, and just weighting them differently according to the number of annotators that identified them. During training, this weight could then be utilized as an additional oversampling factor, thus allowing undisputed annotations to have a greater impact than more doubtful ones. Apart from allowing the classifiers to also accomodate less obvious cases, this procedure would also have the positive effect of generating more training data instances. During testing, the evaluation of resolution decisions could similarly be weighted according to the number of annotators that also made the particular decision, instead of regarding as wrong all resolutions that are not backed by at least, say, three annotators. Note that in the evaluation used in this thesis, a resolution decision by a classifier was regarded as incorrect even if two of the four annotators made the same decision during annotation.

The fact that or corpus consisted of multi-party dialog is another potential source of serious problems, and thus a possible cause of the low resolution performance. Future research could repeat the experiments described in this thesis with a corpus of two-party dialogs annotated according to the principles developed for this thesis. Previous work on two-party dialog (e.g. Byron & Stent (1998), Eckert & Strube (2000), Strube & Müller (2003), Byron (2004)) suggested that the presence of only two speakers avoids some of the referential ambiguities that in our multi-party dialog corpus cause problems for both the annotators and the resolution system. Eckert & Strube (2000) e.g. use what they call the Context Ranking algorithm (Chapter 5.2.3, p. 119) for selecting a clause from which to create a VP antecedent for a given discourse-deictic pronoun. This algorithm crucially depends on clauses to come in *linear* order, i.e. in an order in which clause adjacency can be interpreted in terms of relatedness. In contrast, one of the characteristics of the ICSI Meeting Corpus is that adjacency of utterances in the transcript cannot safely be interpreted in this way. For this thesis, using a different (two-party) dialog corpus was not an option because the corpus to be used was determined beforehand. However, given the fact that none of the features used in this thesis depends on manual annotations, repeating the experiments on a different corpus should be straightforward.

The resolution of anaphors with NP and VP antecedents is fundamentally different from the resolution of anaphors with NP antecedents alone. The need to have available as potential antecedents both NPs and VPs generates a huge number of wrong antecedents for anaphors to choose from. In order to cut down on the number of VP antecedent candidates, a heuristic constraint was applied which generated VP antecedent candidates only for those anaphors which were either *that* or which were the object of constructions involving the verb *do*. A selective qualitative analysis of individual resolution results failed to show a pattern of characteristic properties of anaphor-antecedent pairs which would distinguish ″resolvable″ from ″unresolvable″ ones. For future research, a more detailed error analysis would be necessary. It might also be useful to (at least temporarily) separate the resolution of anaphors with NP and VP antecedents, by manually preselecting those antecedent candidates that are known to be relevant, and suppressing others. While this would result in an unrealistic experimental setting, it would nonetheless probably yield useful insights.

On a more technical level, the rather large number of machine learning features (both standard and novel) and the way they were applied rather indiscriminately might have produced flawed classifiers with corresponding behaviour. Although the individual features (especially the ones that were introduced in this thesis and the ones that were operationalized here for the first time) are all well-motivated and plausible, their practical application in combination with (many) other features might produce unexpected results. An important task for future research would be to perform a quantitative feature analysis of the contributions of individual features. For those classifiers that produce a human-readable model (e.g. J48), a qualitative analysis of the learned model and of the plausibility of the regularities encoded therein would also be very useful.

As a final conclusion, our experimental results bear out the fact that the work described in this thesis only marks the first attempt to fully automatic pronoun resolution in spoken language in general and in spoken multi-party dialog in particular. In this final chapter, we pointed out several ways in which the work that was started with this thesis can and should be continued and improved. In the future, thus, the system described in this thesis will probably at least serve as a reference baseline.

References

Abney, Steven (1996). Partial parsing via finite-state cascades. In *Proceedings of the Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic.*, pp. 8–15.

Agresti, Alan (1990). *Categorical Data Analysis*. New York, N.Y.: Wiley.

Allen, James F. & Lenhart K. Schubert (1991). *The TRAINS Project. TRAINS Technial Note*. Technical Report 91-1: Computer Science Department, University of Rochester.

Aone, Chinatsu & Scott W. Bennett (1995). Evaluating automated and manual acquisition of anaphora resolution strategies. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* Cambridge, Mass., 26–30 June 1995, pp. 122–129.

Arnold, Jennifer E., Maria Fagnano & Michael K. Tanenhaus (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32(1):25–36.

Arnold, Jennifer E., Michael K. Tanenhaus, Rebecca J. Altmann & Maria Fagnano (2004). The old and thee, uh, new. *Psychological Science*, 15(9):578–582.

Artstein, Ron & Massimo Poesio (2005). *Kappa*^{3} *= Alpha (or Beta)*. Technical Report CSM-437: University of Essex Department of Computer Science.

Artstein, Ron & Massimo Poesio (2006). Identifying reference to abstract objects in dialogue. In *Proceedings of the 7th Workshop on the Semantics and Pragmatics of Dialogue,* Potsdam, Germany, 11–13 September 2006, pp. 56–63.

Asher, Nicholas (1993). *Reference to Abstract Objects in Discourse*. Dordrecht, The Netherlands: Kluwer.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto (1999). *Modern Information Retrieval*. New York, NY, Harlow, UK: ACM Press, Pearson Addison-Wesley.

Bagga, Amit & Breck Baldwin (1998). Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28-30 May, 1998, pp. 79–85.

Baldwin, Breck (1997). CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Text, Madrid, Spain, July 1997*, pp. 38–45.

Barwise, Jon (1981). Scenes and other situations. *The Journal of Philosophy*, 77:369–397.

Barwise, Jon & John R. Perry (1983). *Situations and Attitudes*. Cambridge, Mass.: MIT Press.

Bean, David & Ellen Riloff (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the Human Language Technolgy Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May, 2004, pp. 297–304.

Bergsma, Shane & Dekang Lin (2006). Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 33–40. Sydney, Australia: Association for Computational Linguistics.

Biber, Douglas (1992). Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In Jan Svartvik (Ed.), *Directions in Corpus Linguistics*, pp. 213–252. Berlin, New York: Mouton de Gryuter.

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. New York, N.Y.: Springer.

Boyd, Adriane, Whitney Gegg-Harrison & Donna Byron (2005). Identifying non-referential *it*: a machine learning approach incorporating linguistically motivated patterns. In *Proceedings of the ACL Workshop on Feature Selection for Machine Learning in NLP, Ann Arbor, MI, June 2005*, pp. 40–47.

Breiman, Leo, Jerome H. Friedman, Charles J. Stone & R.A. Olshen (1984). *Classification and Regression Trees*. Belmont, Cal.: Wadsworth and Brooks/Cole.

Brennan, Susan E., Marilyn W. Friedman & Carl J. Pollard (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics,* Stanford, Cal., 6–9 July 1987, pp. 155–162.

Briscoe, Ted & John Carroll (1997). Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing,* Washington, D.C., 31 March – 3 April 1997, pp. 356–363.

Buchholz, Sabine (2002). *Memory-based Grammatical Relation Finding*, (Ph.D. thesis). Tilburg, The Netherlands: University of Tilburg.

Byron, Donna K. (2001). The uncommon denominator: A proposal for consistent reporting of pronoun resolution results. *Computational Linguistics*, 27(1):569–577.

Byron, Donna K. (2003). *Annotation of Pronouns and their antecedents. A comparison of two domains*. Technical Report 703: Computer Science Department, University of Rochester.

Byron, Donna K. (2004). *Resolving pronominal reference to abstract entities*, (Ph.D. thesis). Rochester, New York: University of Rochester.

Byron, Donna K. & Amanda Stent (1998). A preliminary model of centering in dialog. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, pp. 1475–1477.

Carletta, Jean (1996). Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.

Carletta, Jean (2006). Announcing the AMI Meeting Corpus. *The ELRA Newsletter*, 11(1):3–5.

Chafe, Wallace (1980). *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production.* Norwood, N.J.: Ablex Publishing Corporation.

Channon, Robert (1980). Anaphoric *that*: a friend in need. In *Papers from the Parasession on Pronouns and Anaphora*, pp. 98–109. Chicago Linguistic Society.

Charniak, Eugene (2000). A maximum-entropy-inspired parser. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics,* Seattle, Wash., 29 April – 3 May, 2000, pp. 132–139.

Chomsky, Noam (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

Christensen, H., Y. Gotoh & S. Renals (2001). Punctuation annotation using statistical prosody models. In *ISCA Workshop on Prosody in Speech Recognition and Understanding.*, pp. 35–40.

Clemente, José Carlos, Kentaro Torisawa & Kenji Satou (2004). Improving the identification of non-anaphoric *it* using Support Vector Machines. In *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications,* Geneva, Switzerland.

Cohen, William W. (1995). Fast effective rule induction. In *Proc. of the 12th International Conference on Machine Learning*, pp. 115–123.

Daelemans, Walter, Jakub Zavrel, Ko van der Sloot & Antal van den Bosch (2004). *TiMBL: Tilburg Memory Based Learner, version 5.1, Reference Guide*. Technical Report ILK 04-02: ILK Tilburg.

Dagan, Ido, John Justenson, Shalom Lappin, Herbert Leass & Ammon Ribak (1995). Syntax and lexical statistics in anaphora resolution. *Applied Artificial Intelligence*, 9(6):633–644.

Dahl, Östen & Christina Hellman (1995). What happens when we use an anaphor? In *Papers from the XVth Scandinavian Conference of Linguistics,* Oslo, Norway, January 13-15, 1995, pp. 79–86.

Di Eugenio, Barbara (2000). On the usage of Kappa to evaluate agreement on coding tasks. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation,* Athens, Greece, 31 May-June 2, 2000, pp. 441–444.

Di Eugenio, Barbara & Michael Glass (2004). The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Dimitrov, Marin, Kalina Bontcheva, Hamish Cunningham & Diana Maynard (2002). A light-weight approach to coreference resolution for named entities in text. In *Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002), Lisbon, Portugal*.

Eckert, Miriam & Michael Strube (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.

Edwards, Jane A. (2003). The transcription of discourse. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (Eds.), *The Handbook of Discourse Analysis*, pp. 321–348. Blackwell Publishing.

Evans, Richard (2001). Applying machine learning toward an automatic classification of *It*. *Literary and Linguistic Computing*, 16(1):45 – 57.

Fellbaum, Christiane (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Fernández, Raquel & Jonathan Ginzburg (2002). Non-sentential utterances: a corpus-based study. *Traitement Automatique des Languages*, 43(2):13–42.

Fleiss, Joseph L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Ginzburg, Jonathan & Ivan Sag (2000). *Interrogative Investigations: The Form, Meaning and Use of English Interrogatives*. Stanford: CSLI.

Godfrey, John J. & Ed Holliman (1993). *Switchboard-1 Transcripts*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia, Penn., USA.

Gross, Derek, James F. Allen & David R. Traum (1993). *The Trains 91 Dialogues. TRAINS Technial Note*. Technical Report 92-1: Computer Science Department, University of Rochester.

Grosz, Barbara, Aravind K. Joshi & Scott Weinstein (1995). Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69:274 – 307.

Gundel, Jeanette K., Michael Hegarty & Kaja Borthen (2003). Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic, Language and Information*, 12(3):281 – 299.

Halliday, M. A. K. & Ruqaiya Hasan (1976). *Cohesion in English*. London: Longman.

Harman, Donna & Mark Liberman (1994). *TIPSTER Complete LDC93T3A*. 3 CD-ROMS. Linguistic Data Consortium, Philadelphia, Penn., USA.

Hays, William L. (1994). *Statistics* (5th ed.). Orlando, FLA: Harcourt Brace & Company.

Heeman, Peter & James Allen (1999). Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue. *Computational Linguistics*, 25(4):527–571.

Heeman, Peter A. & James F. Allen (1995). *The Trains 93 Dialogues. TRAINS Technial Note*. Technical Report 94-2: Computer Science Department, University of Rochester.

Hegarty, Michael, Jeanette K. Gundel & Kaja Borthen (2001). Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27:163–186.

Hirschman, Lynette & Nancy Chinchor (1997). *MUC-7 Coreference Task Definition,* `http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/co_task.html`.

Hirschman, Lynette, Patricia Robinson, John Burger & Marc Vilain (1997). Automating coreference: The role of annotated training data. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, pp. 118–121.

Hobbs, Jerry R. (1978). Resolving pronominal references. *Lingua*, 44:311–338.

Hosmer, David W. Jr. & Stanley Lemeshow (2000). *Applied Logistic Regression* (2nd ed.). John Wiley & Sons, Inc.

Hoste, V. (2005). *Optimization in Machine Learning of Coreference Resolution*, (Ph.D. thesis). University of Antwerp.

Iida, Ryu, Kentaro Inui, Hiroya Takamura & Yuji Matsumoto (2003). Incorporating contextual cues in trainable models for coreference resolution. In *Proceedings of the EACL '03 Workshop on the Computational Treatment of Anaphora,* Budapest, Hungary, 14th April, 2003, pp. 23–30.

Janin, Adam (2002). Meeting recorder. In *Proceedings of the Applied Voice Input/Output Society Conference (AVIOS),* San Jose, California, USA, May 2002.

Janin, Adam, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke & Chuck Wooters (2003). The ICSI Meeting Corpus. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing,* Hong Kong, pp. 364–367.

Japkowicz, Nathalie & Shaju Stephen (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449.

Jurafsky, Daniel & James H. Martin (2000). *Speech and Language Processing*. Upper Saddle River, N.J.: Prentice Hall.

Kabadjov, Mijail A., Massimo Poesio & Josef Steinberger (2005). Task-based evaluation of anaphora resolution: The case of summarization. In *Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research,* Borovets, Bulgaria.

Kaltenböck, Gunther (2005). *It*-extraposition in english: A functional view. *International Journal of Corpus Linguistics*, 10(2):119–159.

Kehler, Andrew (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics*, 23(3):467–475.

Kehler, Andrew, Douglas Appelt, Lara Taylor & Aleksandr Simma (2004). The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technolgy Conference of the North American Chapter of the Association for Computational Linguistics,* Boston, Mass., 2–7 May, 2004, pp. 289–296.

Kertz, Laura, Andrew Kehler & Jeffrey L. Elman (2006). Evaluating a coherence-based model of pronoun interpretation. In *Proceedings of the Workshop on Ambiguity in Anaphora, 18th European Summer School in Logic, Language and Information, Malaga, Spain, August 7-11.*, pp. 49–56.

Krippendorff, Klaus (1980). *Content Analysis: An introduction to its methodology*. Beverly Hills, CA: Sage Publications.

Lapata, Maria, Scott McDonald & Frank Keller (1999). Determinants of adjective-noun plausibility. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics,* Bergen, Norway, 8–12 June 1999, pp. 30–36.

Lappin, Shalom & Herbert J. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Leech, Geoffrey, Paul Rayson & Andrew Wilson (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London, UK: Longman.

Lendvai, Piroska, Antal van den Bosch & Emiel Krahmer (2003). Machine learning for shallow interpretation of user utterances in spoken dialogue systems. In *EACL-03 Workshop on Dialogue Systems: Interaction, Adaptation and Styles of Management,* Budapest, Hungary, pp. 69–78.

Linde, Charlotte (1979). Focus of attention and the choice of pronouns in discourse. In Talmy Givon (Ed.), *Syntax and Semantics, Vol. 12: Discourse and Syntax*, pp. 337–354. Academic Press.

Linguistic Data Consortium (2005). *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 5.6.1.* `http://www.ldc.upenn.edu/Projects/ACE/`.

Luo, Xiaoqiang, Abe Ittycheriah, Hongyan Jing, Nanda Kambhatla & Salim Roukos (2004). A mention-synchronous coreference resolution algorithm based on the Bell Tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics,* Barcelona, Spain, 21–26 July 2004, pp. 136–143.

LuperFoy, Susann (1991). *Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions*, (Ph.D. thesis). University of Texas.

Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Marcus, Mitchell P., Beatrice Santorini & Mary Ann Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics,* 19(2):313–330.

McCarthy, Joseph F. & Wendy G. Lehnert (1995). Using decision trees for coreference resolution. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montréal, Canada, 1995, pp. 1050–1055.

Mel'čuk, Igor A. (2003). Levels of dependency in linguistic description: Concepts and problems. In V. Agel, L. Eichinger, H.-W. Eroms, P. Hellwig, H. J. Heringer & H. Lobin (Eds.), *Dependency and Valency. An International Handbook of Contemporary Research, Vol.1*, pp. 188–229. Berlin, New York: Walter de Gruyter.

Mieskes, Margot (2008). *Exploring Methods for the Automatic Summarization of Meetings,* (Ph.D. thesis). Friedrich-Alexander Universität Erlangen-Nürnberg, Germany. To appear.

Mitchell, Tom M. (1997). *Machine Learning*. McGraw Hill Series in Computer Science. McGraw Hill.

Mitkov, Ruslan (2002). *Anaphora Resolution*. London, UK: Longman.

Moens, Marc & Marc Steedman (1988). Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28.

Mourelatos, Alexander P. D. (1978). Events, processes, and states. *Linguistics and Pilosophy*, 2:415–434.

Müller, Christoph (2006). Automatic detection of nonreferential it in spoken multi-party dialog. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics,* Trento, Italy, 3–7 April 2006, pp. 49–56.

Müller, Christoph, Stefan Rapp & Michael Strube (2002). Applying Co-Training to reference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 352–359.

Müller, Christoph & Michael Strube (2006). Multi-level annotation of linguistic data with MMAX2. In Sabine Braun, Kurt Kohn & Joybrato Mukherjee (Eds.), *Corpus Technology and Language Pedagogy: New Resources, New Tools, New Methods*. Frankfurt a.M., Germany: Peter Lang.

Navarretta, Costanza (2004). Resolving individual and abstract anaphora in texts and dialogues. In *Proceedings of the 20th International Conference on Computational Linguistics,* Geneva, Switzerland, 23 August – 27 August 2004, pp. 233–239.

Ng, Vincent & Claire Cardie (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics,* Philadelphia, Penn., 7–12 July 2002, pp. 104–111.

Ng, Vincent & Claire Cardie (2003). Bootstrapping coreference classifiers with multiple machine learning algorithms. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing,* Sapporo, Japan, 11–12 July 2003, pp. 113–120.

Nicolae, Cristina & Gabriel Nicolae (2006). BestCut: A graph algorithm for coreference resolution. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, 22–23 July, pp. 275–283.

Paice, C. D. & G. D. Husk (1987). Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun 'it'. *Computer Speech and Language*, 2:109–132.

Paice, C. D. & Paul A. Jones (1993). The identification of important concepts in highly structured technical papers. In *SIGIR '93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 69–78. New York, NY, USA: ACM Press.

Passonneau, Rebecca J. (1991). Some facts about centers, indexicals, and demonstratives. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, Cal., 18 – 21 June 1991, pp. 63–70.

Passonneau, Rebecca J. (1994). *Protocol for Coding Discourse Referential Noun Phrases and Their Antecedents*. Technical Report: Columbia University.

Passonneau, Rebecca J. (1997). *Applying Reliability Metrics to Co-Reference Annotation*. Technical Report CUCS-017-97: Columbia University.

Passonneau, Rebecca J. (2004). Computing reliability for co-reference annotation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation,* Lisbon, Portugal, 26-28 May, 2004.

Pfau, Thilo, Don Ellis & Andreas Stolcke (2001). Multispeaker speech activity detection for the icsi meeting recorder. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop,* Madonna di Campiglio, Italy.

Poesio, Massimo & Ron Artstein (2005a). Annotating (anaphoric) ambiguity. In *Proceedings from the Corpus Linguistics Conference*.

Poesio, Massimo & Ron Artstein (2005b). The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pp. 76–83.

Poesio, Massimo & Renata Vieira (1998). A corpus-based investigation of definite description use. *Computational Linguistics*, 24(2):183–216.

Popescu-Belis, Andrei & Isabelle Robba (1998). Three new methods for evaluating coreference resolution. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28-30 May, 1998.

Postal, Paul & Geoffrey Pullum (1988). Expletive noun phrases in subcategorized positions. *Linguistic Inquiry*, 19:635–670.

Preiss, Judita (2002). Anaphora resolution with memory based learning. In *Proceedings of the 5th UK Special Interest Group for Computational Linguistics (CLUK5)*, pp. 1–8.

Quinlan, John R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik (1991). *A Comprehensive Grammar of the English Language* (9th ed.). London, UK: Longman.

Rish, Irina (2001). *An empirical study of the Naive Bayes classifier*. Technical Report RC 22230: IBM Research Division.

Rish, Irina, Joseph H. Hellerstein & Jayram Thathachnar (2001). *An Analysis of Data Characteristics that Affect Naive Bayes Performance*. Technical Report RC 21993: IBM Research Division.

Rocha, Marco (1997). Supporting anaphor resolution in dialogues with a corpus-based probabilistic model. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution,* Madrid, Spain.

Rocha, Marco (1999). Coreference resolution in dialogues in English and Portuguese. In *Proceedings of the ACL Workshop on Coreference Resolution and its applications,* College Park, Maryland, USA.

Schiffman, Rebecca J. (1985). *Discourse constraints on 'it' and 'that': A Study of Language Use in Career Counseling Interviews*, (Ph.D. thesis). Chicago, Illinois: University of Chicago.

Schmid, Helmut (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP). Manchester, U.K., 14–16 September 1994.*

Shriberg, Elizabeth (1994). *Preliminaries to a theory of speech disfluencies*, (Ph.D. thesis). University of California at Berkeley.

Shriberg, Elizabeth, Andreas Stolcke & Don Baron (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01),* Aalborg, Denmark, 3–7 September 2001, Vol. 2, pp. 1359–1362.

Sinclair, John (2004). *Trust the Text: Language, Corpus, and Discourse*. Oxfordshire, UK: Routledge.

Sleator, Daniel & Davy Temperley (1993). Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Steinberger, Josef, Massimo Poesio, Mijail Kabadjov & Karel Jezek (2007). Two uses of anaphora resolution in summarization. *Information Processing and Management: Special Issue on Summarization*, 43(6):1663–1680.

Stolcke, Andreas, Chuck Wooters, Nikki Mirghafori, Tuomo Pirinen, Ivan Bulyko, Dave Gelbart, Martin Graciarena, Scott Otterson, Barbara Peskin & Mari Ostendorf (2004). Progress in meeting recognition: The ICSI-SRI-UW spring 2004 evaluation system. In *NIST ICASSP 2004 Meeting Recognition Workshop,* Montreal, Canada, May 2004.

Stoyanov, Veselin & Claire Cardie (2006). Partially supervised coreference resolution for opinion summarization through structured rule learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sydney, Australia, pp. 336–344.

Strube, Michael (1998). Never look back: An alternative to centering. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics,* Montréal, Québec, Canada, 10–14 August 1998, Vol. 2, pp. 1251–1257.

Strube, Michael & Udo Hahn (1996). Functional centering. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, Cal., 24–27 June 1996, pp. 270–277.

Strube, Michael, Margot Mieskes & Christoph Müller (2007). Gesprächsprotokolle auf Knopfdruck: Die automatische Zusammenfassung von gesprochenen Dialogen. In Werner Kallmeyer & Gisela Zifonun (Eds.), *Sprachkorpora - Datenmengen und Erkenntnisfortschritt*, pp. 249–265. Berlin, New York: de Gruyter.

Strube, Michael & Christoph Müller (2003). A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, 7–12 July 2003, pp. 168–175.

Strube, Michael & Maria Wolters (2000). A probabilistic genre-independent model of pronominalization. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle, Wash., 29 April – 3 May, 2000, pp. 18–25.

Stuckardt, Roland (2003). Coreference-based summarization and question answering: A case for high precision anaphor resolution. In *Proceedings of the 2003 International Symposium on Reference Resolution and Its Application to Question Answering and Summarization (ARQAS), Universita Ca' Foscari, Venice, June 2003*, pp. 33–41.

Tetreault, Joel & James Allen (2004a). Dialogue structure and pronoun resolution. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium*, S. Miquel, Azores, Portugal, September 23–24.

Tetreault, Joel & James Allen (2004b). Semantics, dialogue, and reference resolution. In *Proceedings of the 8th Workshop on the Semantics and Pragmatics of Dialogue*, Barcelona, Spain, July 19–21.

Thompson, Henry S., Anne Anderson, Ellen Gurman Bard, Gwyneth Doherty-Sneddon, Alison Newlands & Cathy Sotillo (1993). The HCRC Map Task corpus: Natural dialogue for speech recognition. In *Proceedings of the HLT Conference Workshop on Human Language Technology*, Princeton, New Jersey, pp. 25–30.

Thompson, Henry S. & David McKelvie (1997). Hyperlink semantics for standoff markup of read-only documents. In *Proceedings of SGML Europe '97*, Barcelona, Spain.

Toutanova, Kristina, Dan Klein & Christopher D. Manning (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL 03*, pp. 252–259.

Trouilleux, François, Éric Gaussier, Gabriel G. Bés & Annie Zaenen (2000). Coreference resolution evaluation based on descriptive specificity. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation,* Athens, Greece, 31 May-June 2, 2000.

van Deemter, Kees & Rodger Kibble (2001). On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics*, 26(4):629–637.

Vendler, Zeno (1967). *Linguistics in Philosophy*. Ithaca, N.Y.: Cornell University Press.

Versley, Yannick (2006). Disagreement dissected: Vagueness as a source of ambiguity in nominal (co-)reference. In *Proceedings of the Workshop on Ambiguity in Anaphora, 18th European Summer School in Logic, Language and Information, Malaga, Spain, August 7-11.*, pp. 83–89.

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52. San Mateo, Cal.: Morgan Kaufmann.

Walker, Marilyn A., Masayo Iida & Sharon Cote (1994). Japanese discourse and the process of centering. *Computational Linguistics*, 20(2):193–233.

Webber, Bonnie L. (1991). Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Witten, Ian H. & Eibe Frank (1999). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco: Morgan Kaufman.

Yang, Xiaofeng (2005). *A twin-candidate model for learning based coreference resolution*, (Ph.D. thesis). Singapore: National University of Singapore.

Yang, Xiaofeng, Jian Su & Chew Lim Tan (2005). Improving pronoun resolution using statistics-based semantic compatibility information. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics,* Ann Arbor, Mi., USA, 25–30 June 2005, pp. 165–172.

Yang, Xiaofeng, Guodong Zhou, Jian Su & Chew Lim Tan (2003). Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics,* Sapporo, Japan, 7–12 July 2003, pp. 176–183.

Yeh, Jui-Feng & Chung-Hsien Wu (2006). Edit disfluency detection and correction using a cleanup language model and an alignment model. *IEEE Transactions on Audio, Speech and Language Processing*, 14(5):1574–1583.