

Discourse-Givenness of Noun Phrases Theoretical and Computational Models

Dissertation

zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)
der Humanwissenschaftlichen Fakultät
der Universität Potsdam

vorgelegt von
Julia Ritz

Stuttgart, Mai 2013

Published online at the
Institutional Repository of the University of Potsdam:
URL <http://opus.kobv.de/ubp/volltexte/2014/7081/>
URN [urn:nbn:de:kobv:517-opus-70818](http://nbn-resolving.org/urn:nbn:de:kobv:517-opus-70818)
<http://nbn-resolving.de/urn:nbn:de:kobv:517-opus-70818>

Erklärung (*Declaration of Authenticity of Work*)

Hiermit erkläre ich, dass ich die beigefügte Dissertation selbstständig und ohne die unzulässige Hilfe Dritter verfasst habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel genutzt. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Ich versichere außerdem, dass ich die Dissertation in der gegenwärtigen oder einer anderen Fassung nur in diesem und keinem anderen Promotionsverfahren eingereicht habe, und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Im Falle des erfolgreichen Abschlusses des Promotionsverfahrens erlaube ich der Fakultät, die angefügten Zusammenfassungen zu veröffentlichen.

Ort, Datum Unterschrift

Contents

Preface	xiii
1 Introduction	1
1.1 Discourse-Givenness and Related Concepts	3
1.2 Goal	4
1.3 Motivation	6
1.4 The Field of Discourse-Givenness Classification	10
1.5 Structure of this Document	11
2 Concepts and Definitions	13
2.1 Basic Concepts and Representation	13
2.1.1 Basic Concepts in Discourse	14
2.1.2 Formal Representation	15
2.2 Reference	18
2.2.1 Identifying Basic Units	20
2.2.2 Non-Referring Use of Expressions	24
2.2.3 World of Reference	25
2.2.4 Specificity	27
2.2.5 Generalizations	29
2.2.6 Vagueness and Ambiguity	33
2.3 Coreference	35
2.3.1 Consequences of Vagueness, Ambiguity and Non-Specificity	36
2.3.2 Coreference vs. Lexical, Intensional or Extensional Identity	37
2.3.3 Identity of Binding Index	39
2.3.4 Asserted Identity	42
2.3.5 Identity and Time Dependence	45
2.3.6 Accommodation of Referents	45
2.3.7 Bridging	46
2.4 Coreference, Discourse-Givenness and Information Status	48
3 Corpora and Annotation Schemes	53
3.1 Corpora Annotated with Discourse-Givenness or Coreference	53
3.1.1 MUC-7	54
3.1.2 Nissim’s Annotation of the Switchboard Corpus	55
3.1.3 OntoNotes	55
3.1.4 ARRAU	55

3.1.5	PCC	56
3.1.6	TüBa-D/Z	56
3.1.7	DIRNDL	57
3.2	Comparison of Annotation Schemes	57
3.2.1	Markables and Preconditions	58
3.2.1.1	Markables	58
3.2.1.2	Preconditions	68
3.2.2	Coreference and Context	69
3.2.2.1	Discourse-Givenness and Coreference	69
3.2.2.2	Notions of Context	74
3.2.3	Formalization and Evaluation of the Annotation	74
3.2.3.1	Formalization	74
3.2.3.2	Evaluation	75
3.2.4	A Critical Assessment of Existing Corpora	77
4	Related Work and Technical Background	81
4.1	Data and Categories	81
4.2	Features	83
4.2.1	Properties of NPs	83
4.2.2	NPs' Relations to their Context	85
4.3	Algorithms	87
4.3.1	Evaluation Measures	93
4.4	Experiments and Results	95
5	Discourse-Givenness Classification: New Experiments and Results	101
5.1	Data	101
5.2	Methods	102
5.2.1	Features	102
5.2.2	Algorithms and Evaluation Measures	109
5.3	Classification Results	110
5.3.1	Quantitative Results	110
5.3.1.1	OntoNotes	110
5.3.1.2	MUC-7	116
5.3.1.3	ARRAU	123
5.3.1.4	TüBa-D/Z	126
5.3.2	Discussion	132
5.3.2.1	Machine Learning Aspects	132
5.3.2.2	Linguistic Aspects	134
5.4	Comparison to Related Work	135
5.4.1	OntoNotes	136
5.4.2	MUC-7	136
5.4.3	Related Work in General	136
6	Conclusions and Outlook	139
6.1	Conclusions	139
6.2	Outlook	140

7 Summaries	143
7.1 Summary in English	143
7.2 Summary in German (<i>Zusammenfassung in deutscher Sprache</i>)	146
References	150
Index	166

List of Figures

1	Coreference visualisation in ANNIS2	xiv
1.1	Example Discourse	1
1.2	Example Structure according to Rhetorical Structure Theory (RST)	9
1.3	Coreference Resolution: Terminology	10
2.1	DRS of ‘Jones owns Ulysses. It fascinates him.’	16
2.2	DRS of ‘Susan has found every book which Bill needs. They are on his desk.’	17
2.3	DRS of ‘Most students bought books that would keep them fully occupied during the next two weeks.’	17
2.4	DRS of ‘Mary wants to marry a rich man. He is a banker.’	26
2.5	DRS of ‘Mary wants to marry a rich man. He must be a banker.’	26
2.6	DRS of ‘If parents are dissatisfied with a school, they should have the option of switching to another.’	27
2.7	DRS of ‘Parents should have the option of switching to another school, if they are dissatisfied with the current school.’	27
2.8	DRS of ‘Jones does not own a Porsche.’	30
2.9	DRS of ‘A wolf takes a mate for life.’	30
2.10	DRS of ‘Sheep are quadrupedal, ruminant mammals typically kept as livestock. Sheep are members of the order Artiodactyla.’	33
2.11	Syntactic Tree Containing Traces	42
2.12	DRS for ‘John took Mary to Acapulco. They had a lousy time.’	46
2.13	DRS for ‘The students went camping. They enjoyed it.’	47
2.14	Prince’s (1981) Hierarchical Scheme of Familiarity	49
2.15	Information Status, Discourse Status, and Hearer Status	50
3.1	Syntactic Analysis of Possessives in MUC-7	60
3.2	Syntactic Analysis of Possessives in OntoNotes 1.0	60
3.3	Syntactic analysis of a PP with Fusion of Preposition and Definite Determiner	63
4.1	Example Decision Tree for Discourse-Givenness.	88
4.2	Pseudo-code for Decision Tree Construction	88
4.3	Example Rules for Discourse-Givenness.	89
4.4	Pseudo-code for Rule Learners	89
4.5	Support Vector Machines	91
5.1	OntoNotes 1.0: Decision Tree	113

5.2	OntoNotes 1.0: Learning Curve	114
5.3	MUC-7: Decision Tree	118
5.4	MUC-7: Learning Curve (original split)	120
5.5	MUC-7: Learning Curve (random split)	122
5.6	MUC-7: Learning Curve (all vs. Formaleval)	122
5.7	ARRAU 1.2: Learning Curve	124
5.8	ARRAU 1.2: Decision Tree	125
5.9	TüBa-D/Z 6.0: Learning Curve	129
5.10	TüBa-D/Z 6.0: Decision Tree (coreferential/anaphoric/bound)	130
5.11	TüBa-D/Z 6.0: Decision Tree (coreferential)	131

List of Tables

2.1	Types of Reference (Tentative)	19
2.2	Specificity, Genericity and Coreference	32
2.3	Concepts and Definitions	51
3.1	Overview of Corpora Annotated with Coreference/Discourse-Givenness	54
3.2	Markable Definitions (part I: English Corpora)	64
3.2	Markable Definitions (part II: German Corpora)	66
3.3	Definitions of Coreference (part I: English Corpora)	70
3.3	Definitions of Coreference (part II: German Corpora)	72
4.1	Overview: Approaches to Identifying Information Structure	82
4.2	Comparison of Learning Algorithms	92
4.3	Naming Conventions for Evaluation	93
4.4	Example Contingency Table	95
4.5	State of the Art Classification Results (part I)	96
4.5	State of the Art Classification Results (part II)	97
5.1	Feature ‘maxsimilar_mention’: Calculation Example	105
5.2	Features Used in the Classification Experiments	109
5.3	OntoNotes 1.0: Class Distribution	111
5.4	OntoNotes 1.0: Class Distribution in Training, Development and Test Set (random split)	111
5.5	OntoNotes 1.0: Classification Results	112
5.6	OntoNotes 1.0: Classification Results (different algorithms)	114
5.7	MUC-7: Class Distribution	116
5.8	MUC-7: Class Distribution in Train, Dryrun and Formaleval Set (original split)	116
5.9	MUC-7: Proportions of Pronouns in Different Sets	117
5.10	MUC-7: Classification Results (original split), part I	119
5.10	MUC-7: Classification Results (original split), part II	120
5.11	MUC-7: Classification Results (5-fold cross-validation, hold-out)	121
5.12	ARRAU 1.2: Class Distribution	123
5.13	ARRAU 1.2: Class Distribution in Training, Development and Test Set (random split)	123
5.14	ARRAU 1.2: Classification Results (random split, all NPs vs. referring NPs)	124
5.15	TüBa-D/Z 6.0: Class Distribution	126

5.16	TüBa-D/Z 6.0: Class Distribution in Training, Development and Test Set (random split)	127
5.17	TüBa-D/Z 6.0: Classification Results	128
5.18	OntoNotes 1.0: κ values for classifiers vs. original annotation	133
5.19	MUC-7: κ values for classifiers vs. original annotation	133
5.20	ARRAU 1.2: κ values for classifiers vs. original annotation	133
5.21	TüBa-D/Z 6.0: κ values for classifiers vs. original annotation	134
5.22	MUC-7: Comparison of Classification Results	137

Preface

Acknowledgements

Having completed this thesis, I have to admit that the most important things I have learnt are not documented in it.

First of all, I thank my advisors Stefan Evert and Manfred Stede for their technical and organisational advice, readiness to discuss ideas, and general support. Also, I thank the members of the commission.

I am grateful for having met such a lot of researchers at Potsdam University who share an interest in Linguistics and Natural Language Processing (and coffee). In particular, these were my fellow PhD students from the PhD colloquium, Timo Baumann, Okko Buss, Konstantina Garoufi, Peter Kolb, Florian Kuhn, and Andreas Peldszus, as well as my colleagues Christian Chiarcos, Stefanie Dipper, Michael Götze, Amir Zeldes, and Florian Zipser. Thank you for giving your views, asking questions, sharing your knowledge and information, and for listening.

Thanks for inspiring papers, discussions and reviews to Heike Bieler, Gerlof Bouma, Halyna und Jan Finzen, Aurélie Herbelot, Alexander Koller, Jonas Kuhn, Stavros Skopeteas, Malte Zimmermann and Heike Zinsmeister. Thanks to Ruben van de Vijver for his advice to “write every day”.

I am indebted to Massimo Poesio and Arndt Riestler for their providing me with their corpora ARRAU and DIRNDL, as well as Aoife Cahill and Kerstin Eckart for additional information on corpora and classification studies.

I am also indebted to John Gill and my family for proofreading.

Finally, I thank from the heart my friends and family for their patience, understanding, and constant support.

Much has been said on the topics of reference, coreference, and discourse-givenness. A most concise illustration of challenges to the definition of these concepts can be found in literature:

“I love talking about nothing [...]. It is the only thing I know anything about.”
(Oscar Wilde, *An Ideal Husband*)

About This Document

Preliminary results of this work have been published in Ritz (2010) (Sections 5.2.1 and 5.4.1), and an excerpt of related work in Lüdeling et al. (to appear). An earlier version of Chapter 2 has been used for teaching purposes by Stefan Evert and Stefanie Dipper. The corpus search and visualisation tool ANNIS2 was used to inspect the data in context and test hypotheses on the effectiveness of certain features. A description of the ANNIS2 framework can be found in Zeldes et al. (2009), and a description of the procedure of error analysis with ANNIS2 in Chiarcos and Ritz (2010).

In ANNIS2, search matches are highlighted in red. Here, this is the expression ‘*Messrs. Lee and Bynoe*’ and its antecedent ‘*the partners*’. Coreferent expressions are underlined in the same colour. On clicking on an expression, e.g. ‘*Lee*’, the expression and all its coreferent expressions appear shaded in the same colour (blue or purple, respectively, in the figure below).

Chicago businessmen Bertram M. Lee and Peter Bynoe signed a new agreement to purchase the Denver Nuggets basketball team, but not as principal owners. On Saturday, the partners said 0 the team would be purchased for \$ 54 million by a new group including Comsat Video Enterprises Inc., a unit of Communications Satellite Corp. based here. Comsat Video will pay \$ 17 million for a 62.5 % interest, with Messrs. Lee and Bynoe putting up \$ 8 million for a 37.5 % stake in the team. Under terms of the sale, Nuggets owner Sidney Shlenker could receive up to \$ 11 million in additional payments from the franchise's future earnings. Messrs. Lee and Bynoe last July announced a deal that would have made them the first black principal owners of a major professional sports franchise. But the deal fell apart last week for lack of financing. Comsat Video is headed by Robert Wussler, who resigned his No. 2 executive post with Turner Broadcasting System Inc. just two weeks ago to take the Comsat position.

Figure 1: Coreference visualisation in ANNIS2: Classification error ‘*Messrs. Lee and Bynoe*’ in its context

Examples have been chosen with respect to their linguistic significance; their content does not represent views of the author. In some of the explanatory examples, annotation from the original corpora has been simplified to some extent, leaving out irrelevant parts for reasons of readability. Textbook examples cited from the literature may be provided with additional or altered markup for the purpose of a uniform presentation across this work.

List of abbreviations, symbols and technical terms

The expressions listed below are assumed familiar (as are words like *mention*, or *to combine*, which maintain their original meaning when used as technical terms). All other relevant concepts will be explained and can be found using the index at the end of this volume.

#	the following utterance (or part of utterance) is infelicitous.
?	the annotation or following (part of) utterance is questionable.
anaphor	1. phrase referring back to another phrase 2. phrase only interpretable with the help of the preceding context 3. stylistic device (repeatedly starting sentences using the same word or phrase).
annotation	additional information anchored to (parts of) a text, e.g. parts of speech for each word
antecedent	a referent's earlier mention in the same text
baseline	point of reference performance for the purpose of comparison. Usually produced by a simpler or more shallow system
binding	relation, e.g. between a pronoun and its antecedent
cataphor	pronomial first mention of a referent (before any nominal mention), e.g. <i>When <u>he</u> retires, Al hopes to travel to New Zealand.</i>
collocation	recurring word combination with non-compositional meaning. Further criteria: non-modifiability and non-substituability of the words it consists of (Manning and Schütze, 1999). E.g. <i>to pay attention</i> (better than <i>to give attention</i>), <i>strong tea</i> (not <i>intense tea</i>)
corpus	large collection of (linguistically annotated) texts
denote	to signify. A word/phrase denotes all objects compliant with its meaning (the phrase <i>green bottles</i> denotes all green bottles that have been and will be in the world)
feature, binary	(also: boolean) taking one of two possible values: 1 or 0 (1 is typically interpreted as <i>yes</i> , 0 as <i>no</i>)
categorical	having an enumerable, predefined set of mutually exclusive values
idiom	word or phrase with figurative meaning. Prototypical example: <i>to kick the bucket</i> . Criteria: translation test (cannot be translated literally), its meaning is not obvious to a non-native speaker (though some languages have analogous expressions)
illocutionary act	the use of an utterance to a certain consequence (e.g. an instruction, a promise, etc.)
iff	if and only if
NP	noun phrase
operator	an element representing an instruction, e.g. <i>and</i> for the combination of two truth values
prosody	melodic and rhythmic component of speech
quantifier	a symbol representing a generalization (e.g. the universal quantifier \forall 'for all')
relation, unary	relation taking one argument
binary	taking two arguments
string matching	method for comparing sequences of letters
variable	a symbol used in formal representations

Chapter 1


Introduction

A large part of the world's information is exchanged in the form of natural language (Manning et al., 2009). Human readers have a capacity for text understanding, i.e. they are able to access the information contained in language data.

Consider the text in Figure 1.1¹, in particular the last sentence in the second paragraph.

Berlin

From Wikipedia, the free encyclopedia

Coordinates:  52°30′2″N 13°23′56″E

This article is about the capital of Germany. For other uses, see [Berlin \(disambiguation\)](#).

Berlin (English pronunciation: /bɜːrˈlɪn/; German pronunciation: [bɛʁˈliːn] (listen)) is the capital city of Germany and is one of the 16 states of Germany. It has a population of 3.4 million people,^[1] and is Germany's largest city. It is the second most populous city proper and the eighth most populous urban area in the European Union.^[3] Located in northeastern Germany, it is the center of the Berlin/Brandenburg Metropolitan Region, comprising 4.4 million people from over 190 nations.^[4] Geographically embedded in the European Plains, Berlin is influenced by a temperate seasonal climate. Around one third of the city's territory is composed of forests, parks, gardens, rivers and lakes.^[5]

First documented in the 13th century, Berlin was successively the capital of the Kingdom of Prussia (1701–1918), the German Empire (1871–1918), the Weimar Republic (1919–1933) and the Third Reich (1933–1945).^[6] Berlin in the 1920s was the third largest municipality in the world.^[7] After World War II, the city was divided; East Berlin became the capital of East Germany while West Berlin became a de facto West German exclave, surrounded by the Berlin Wall (1961–1989).^[8] Following German reunification in 1990, the city regained its status as the capital of all Germany hosting 147 foreign embassies.^{[9][10]}

Berlin is a world city of culture, politics, media, and science.^{[11][12][13][14]} Its



Figure 1.1: Example Discourse

When processing this sentence, we segment it into words, then form constituents (e.g. *German reunification*, *the city*), and identify the verb's arguments (subject *the city*, object *its status as the capital of all Germany*). We recognize that the sentence says something about one particular city. As there are many cities in the world, we use

¹Source of text: <http://en.wikipedia.org/wiki/Berlin>, last access February 22th, 2012.

the text preceding the noun phrase *the city* (underlined in red) to infer which city the author relates to. In other words, we categorize the noun phrase as *discourse-given*, i.e. referring to something already mentioned, and establish a relation to that entity already mentioned. In this case, *the city* refers to the same entity in the real world as *Berlin*.

Finally, we retrieve the fact that Berlin became the capital of Germany again after the reunification in 1990. Thus, determining the *discourse-givenness* of an expression is one important step in text understanding.

With vast amounts of language data available, there is an increasing need for Natural Language Processing (NLP), i.e. automated methods and tools that allow to access and exploit the informational content of this data in an efficient way. NLP tools have been developed for many specialized tasks, including the following:

- **Information Extraction:** in information extraction, facts relevant to a certain domain (e.g. changes in the management of companies etc.) are extracted and represented in a more uniform or concise way (tables, timelines, generated texts etc.).
- **Question Answering:** a question answering system accepts user-defined questions, extracts relevant facts and generates answers.
- **Summarization:** summarization systems, when provided with one or more texts, produce a summary, i.e. a text that is shorter than the original text(s) while retaining the most important information.
- **Machine Translation:** machine translation systems translate a text in the source language into a text in the target language.
- **Textual Entailment:** a system for textual entailment checks whether for two fragments of text, from the facts in the first, the facts in the second can be inferred
- **Content Assessment:** systems for content assessment score texts (e.g. students' essays), rating textual cohesion etc.

All of these computational applications produce and/or process structured representations – usually tables or graphs – of what a discourse (e.g. a written text or a dialogue) conveys. This involves

1. identifying referring expressions, i.e. expressions which correspond to objects in the real world (e.g. to the city of Berlin in the example above),
2. determining relations between those objects that are expressed in sentences (e.g. *Berlin is the capital city of Germany*),
3. coreference resolution, i.e. a grouping together of all expressions that refer to the same object (e.g. *Berlin, It, the city* etc.), and
4. accumulating all the information given on a certain object (*Berlin is the capital city of Germany, and has 3.4 million inhabitants, etc.*).

Coreference Resolution (step 3) is crucial to the accumulation of information (step 4). At the same time, it is one of the most difficult steps in processing, both for human readers (cf. Poesio and Artstein (2005) for English and Versley (2006) for German) and for computers. Sometimes, there are even several resolution possibilities.

To solve this complex task, a decomposition into two subtasks has been suggested:²

1. identify those expressions that are *discourse-given*, i.e. referring back to something mentioned earlier;
2. locate the corresponding expressions referred back to.

In this way, a considerable number of NLP tools can benefit from an automatic detection of *discourse-given* expressions.

1.1 Discourse-Giverness and Related Concepts

Within a text or discourse, we can use different linguistic entities to refer to the same object (*referent*) in the real world. An expression is *discourse-given* if its referent has been mentioned in the previous context. In Example (1)³ below, all discourse-given expressions referring to the same object (here: an organization) as *USACafes Limited Partnership* are put in bold face.

A previous mention of the same referent is called an antecedent: *USACafes Limited Partnership* is the antecedent of *it*; *USACafes Limited Partnership* and *it* are antecedents of *its* etc.

The first mention of an object (*USACafes Limited Partnership* in the example) is not discourse-given, it is *coreferent*: there are expressions in the text referring to the same object, but not in the expression's previous context. In the example, all expressions that are coreferent with *USACafes Limited Partnership* are underlined.

- (1) USACafes Limited Partnership said **it** completed the sale of **its** Bonanza restaurant franchise system to a subsidiary of Metromedia Co. for \$ 71 million in cash. **USACafes**, which is nearly half-owned by Sam and Charles Wyly of Dallas, said **it** will distribute proceeds from the sale to unit holders as a liquidating dividend as soon as possible. The Bonanza franchise system, which generates about \$ 600 million in sales annually, represented substantially all of **the partnership's** assets. The sale of the system has been challenged in a class-action suit on behalf of unit holders filed last week in a Delaware court, **USACafes** said. **The company** said **it** believes the suit is without merit.

While discourse-giverness prediction rates an expression as either *discourse-given* or *new*, coreference resolution identifies which expressions refer to the same entities: it forms one group of all expressions referring to *USACafes*, another group of all expressions referring to *USACafe's Bonanza franchise system*, etc. (see Example (2)). Throughout this dissertation, subscript indices and underlining will be used. Identical indices within a

²Other decomposition approaches have been suggested; these are discussed in Section 1.4.

³All examples in this chapter are adapted from the OntoNotes corpus (Hovy et al., 2006) unless specified otherwise.

text represent coreference. The underlining indicates the constituent the index is assigned to.

- (2) USACafes Limited Partnership₁ said it₁ completed the sale of its₁ Bonanza restaurant franchise system₂ to a subsidiary of Metromedia Co. for \$ 71 million in cash₃. USACafes₁, which is nearly half-owned by Sam and Charles Wyly of Dallas, said it₁ will distribute proceeds from the sale₃ to unit holders as a liquidating dividend as soon as possible. The Bonanza franchise system₂, which generates about \$ 600 million in sales annually, represented substantially all of the partnership's₁ assets. The sale of the system_{2,3} has been challenged in a class-action suit₄ on behalf of unit holders filed last week in a Delaware court, USACafes₁ said. The company₁ said it₁ believes the suit₄ is without merit.

Other terms for the phenomenon of discourse-givenness are discourse status (*given/old* vs. *new information*, cf. Prince (1981; 1992)) and anaphoricity (*anaphoric*⁴ vs. *non-anaphoric noun phrases*, cf. Ng and Cardie (2002), Ng (2004), Denis and Baldrige (2007)). A related concept is information status (also called cognitive status), which comprises categories for expressions that are neither given in the discourse nor new to the recipient. Different distinctions have been proposed (Prince, 1981; Prince, 1992; Gundel et al., 1993; Calhoun et al., 2005; Götze et al., 2007; Riester et al., 2010). An overview of these distinctions is given in Section 2.4.

1.2 Goal

The goal of this dissertation is to build classifiers for English and German that determine for any noun phrase (NP) whether or not it is *discourse-given*, i. e. referring back to some entity mentioned earlier in the same text. Full anaphora or coreference resolution is beyond the scope of this work. I consider the task of discourse-givenness classification for NPs a clearly delimitable task and the resulting classification component a useful, reusable resource (see Section 1.3). At the same time, it represents a step in the direction of coreference resolution. Anaphora and coreference resolution differ from discourse-givenness classification in that they require different strategies for the identification of antecedent relations as well as for the evaluation. Antecedent identification requires the comparison of an anaphor to either a set of potential antecedents, or alternatively to each potential antecedent. This might involve, for instance, the testing of syntactic constraints like *c-command*⁵ etc. As for the evaluation, a range of evaluation measures have been proposed, with different methods for taking partially correct assignments to coreference sets into account, and the debate is ongoing.

As human judgement is needed throughout the development process and the evaluation, the languages were chosen based on the language skills of the author.

⁴Generally, the term *anaphor* is used for pronouns or for expressions that are not interpretable without their context. Here, it is used in its broader sense, including all expressions having an antecedent that is coreferent or binds the anaphor.

⁵“Node A c(onstituent)-commands node B if neither A nor B dominates the other[,] and the first branching node which dominates A dominates B.” (Reinhart, 1976, p. 32)

Classifiers need large amounts of data for training. To the best of my knowledge, there is no publicly and/or freely available corpus annotated with discourse-givenness, information status or a similar concept. For this reason, I use coreference annotation to deduce the status of each NP, *discourse-given* or *discourse-new*⁶.

The corpora annotated for coreference available for English and German include (see Poesio and Artstein (2008), Ng (2010) and Pradhan et al. (2011) for overviews):

- MUC-6 and MUC-7 (originating from the Message Understanding Conferences, Chinchor and Sundheim (2003), Chinchor (2001)) and its successor ACE (Automatic Content Extraction, Doddington et al. (2000))⁷
- OntoNotes (Weischedel et al., 2007)
- ARRAU (Anaphora Resolution and Underspecification, Poesio and Artstein (2008)) and its predecessors GNOME (Generating Nominal Expressions, Poesio (2004a)), MATE (Multilevel Annotation Tools Engineering, Poesio et al. (1999), Poesio (2004c)), and the so-called Vieira-Poesio corpus (Poesio and Vieira, 1998)
- PCC (Potsdam Commentary Corpus, (Stede, 2004))
- TüBa-D/Z (Tübinger Baubank des Deutschen/Zeitungskorpus, ‘Tübingen Treebank of German/Newswire Section’, Telljohann et al. (2003), Naumann (2006))
- DIRNDL (Discourse Information Radio News Database for Linguistic analysis, Eckart et al. (2012))

These corpora implement different definitions of coreference. On the basis of a theoretic concept of coreference (Chapter 2), these different definitions are compared in Chapter 3, and several classifiers are trained (Chapter 5). MUC-7, OntoNotes, ARRAU and TüBa-D/Z are used for this purpose.

Despite considerable research effort, current models for the classification of discourse-givenness still have substantial potential for improvement to the point where they are applicable for automated corpus annotation or in NLP systems like Coreference Resolution or Text-to-Speech (TTS) systems.⁸ Thus, one task is to implement new features that help distinguish discourse-given NPs. In particular, this is done by applying new, fuzzier strategies for comparing NPs, as well as making more use of the NP’s immediate context. In Example (3), the repeated use of the verb *sell* (in italics) can be used to help establish a coreference relation between the NPs *four savings-and-loan institutions* and *The four S&Ls*.⁹

⁶The class *discourse-new* contains all NPs that are not *discourse-given*; thus, non-referring NPs are included in the class *discourse-new* (see Sections 2.2.2 and 3.2.1.2 for discussion), as are NPs referring to an object that is mentioned only once (so-called singletons) and NPs that introduce a referent (so-called first mentions).

⁷Annotation in this corpus is limited to expressions referring to persons, organizations, locations, facilities, weapons, vehicles, and geo-political entities.

⁸Previous work is discussed in Section 4.4 in more detail.

⁹Cf. Mitkov’s (1999) syntactic and semantic parallelism.

- (3) The government *sold* the deposits of four savings-and-loan institutions₁, in its first wave of sales of big, sick thrifts, but low bids prevented the sale of a fifth. The four S&Ls₁ *were sold* to large banks, as was the case with most of the 28 previous transactions [...].

1.3 Motivation

A classifier for discourse-givenness can be applied in various NLP systems, e.g.

- as a preprocessing filtering step in Coreference Resolution,
- to enhance prosody (mainly intonation and stress) in Text-to-Speech systems, and
- forming part of an automatic Information Structure and Discourse Analysis

The options and benefits are sketched in the following.

Coreference Resolution

Coreference resolution systems group together expressions that have the same referent. It can be broken down into two subtasks:

- (i) filtering out non-anaphoric noun phrases, and then
- (ii) identifying the antecedents of the remaining entities.

It has been claimed that ruling out non-anaphoric NPs helps limiting the search space and can thus improve a system’s quality and performance (Harabagiu et al., 2001; Ng and Cardie, 2002; Elsner and Charniak, 2007; Uryupina, 2009; Rahman and Ng, 2009; Ng, 2010; Poesio et al., 2004; Kabadjov, 2007, the last two refer to definite descriptions)¹⁰. Recent approaches have re-integrated the two subtasks again: Denis and Baldrige (2007) and Rahman and Ng (2009), for instance, train joint models.

While this work does not extend to Coreference Resolution itself, its outcomes are of relevance to solving the task: methods and features which improve the classification of discourse-givenness are likely to yield improvements in Coreference Resolution, independently of a separate or joint modelling.

As mentioned above, Coreference Resolution is crucial in various NLP tools, including Information Extraction, Question Answering, Summarization and many more.

Text to Speech (TTS)

Text-to-Speech (or speech synthesis) systems read out text. TTS systems are of service to persons with impaired vision or speech. They can also help in situations where the visual attention of the user is needed elsewhere: for instance, they can read out instructions to a person repairing a complex device.

¹⁰Poesio (2004b) defines definite descriptions as extending to “proper names, ‘the’-[NPs], ‘this’-[NPs], ‘that’-[NPs], pronouns, and possessive NPs” (p. 2).

Current TTS systems are reported to sound “unnatural” (Rashad et al., 2010, p. 87), “monotonous” (Bonafonte et al., 2009, p. 131), and with “poorer clarity and prosody of synthesis relative to natural speech” (Stevens et al., 2005, p. 130). Prosodic cues like intonation, stress and rhythm help hearers understand an utterance.

Prosodic deficits in existing TTS systems could be overcome based on observations that relate a constituent’s information structural properties to its prosodic realization. For English, Chafe (1970) and Brown (1983) found that expressions representing given information receive “low pitch” (Brown, 1983, p. 68).¹¹ Umbach (2003) shows that a noun phrase without an accent is interpreted as *given*, whereas it is interpreted as *new*¹² if it is accented, see Example (4) (taken from Umbach (2003), p. 313 capital letters are used to represent accent).

- (4) (John has an old cottage₁.)
- a. Last summer he RECONSTRUCTED the shed₁.
 - b. Last summer he reconstructed the SHED₂.

Similar patterns have been observed for German: Féry and Kügler (2008) find that “givenness lowers [tones] in prenuclear position and cancels them out postnuclearly” (p. 681). Baumann and Riester (2012) show that “[g]iven referents [...] encoded by given lexical items [...] are deaccented” (Baumann and Riester, 2012, p. 137, 139) in read speech.

This evidence leads to conclude that using information on the discourse-givenness of constituents for modelling pitch accent can increase the intelligibility and ‘naturalness’ (i.e. closeness to the prosody of a human reader) of utterances.

Hiyakumoto et al. (1997), Cassell et al. (2001) have used givenness and theme/rheme analyses of sentences to generate an appropriate intonation of these sentences. Albrecht et al. (2005) use givenness and contrast. The respective systems have only been evaluated qualitatively and on a general scale. The influence of givenness as a feature has not been investigated. This may be one reason that givenness is not used at a larger scale.

Information Structure and Discourse Analysis

Understanding a discourse incorporates comprehending its structure. An explicit structure analysis can be used for further processing of the discourse, e.g. for creating excerpts or summaries, or for selecting the documents which are most relevant to a certain user query in Information Retrieval. A tool for the automatic analysis of discourses with respect to information structure and rhetorical structure would therefore be a valuable resource.

The term information structure relates to the ‘packaging’ (Chafe, 1976) and ordering of the content one wants to convey when making an utterance or writing a text. Structuring aims at optimizing the utterance or text in a way that makes the message easily

¹¹Besides givenness, other factors may play a role like surface position and the “persistence of grammatical function” (Hirschberg and Terken, 1993, p. 1362). Nakatani (1996) gives an overview of studies that relate givenness/newness and accent.

¹²The distinction between *discourse-new* and *hearer-new* will only be made from Section 2.4.

understandable to the recipient. This structuring can take place at sentence level and text level.¹³

According to Götze et al. (2007), which is based on Krifka’s (2008) basic notions of information structure, three information structural layers can be distinguished:

- (i) information status (which comprises discourse-givenness)
- (ii) topic and comment
- (iii) focus and background

Information status categorizes the “retrievability” (Götze et al. (2007), p. 150) of a referent either as *given*, *accessible* (the referent is related to a referent in the discourse, and “the inference relation is shared between speaker and hearer” (Götze et al., 2007, p. 150)) or else *new*.

A *topic* is what a sentence is about. The term is used for the referring expression as well as for the referent (Krifka, 2008). Statements are divided into topic and *comment* (what is said about the topic). Expressions that restrict the main predication (e.g. adverbials like *financially* in the sentence ‘*Financially, the Joneses are fine.*’) are called *frame-setters*. For illustrating the term *topic* and how it relates to the mental storage of information, the metaphor of a file-card like system (Reinhart, 1982) is often used: the topic of a sentence provides the recipient with a location (*‘anchor’*) where a piece of information should be stored.

The *new information focus* is the information which develops the discourse and brings it forward (stating the unfamiliar or unexpected). Less relevant information forms the *background*. According to Götze et al. (2007), “focus on a subexpression indicates that it is selected from possible alternatives that are either implicit or given explicitly, whereas the background can be derived from the context of the utterance” (p. 170).

These categories of topic and focus are by definition related to, but not based on discourse-givenness. Often, a sentence’s topic is a given NP, about which the sentence makes a new predication (cf. the discussions in Chafe (1976) and Gundel (1988) among others). This suggests that discourse-givenness, in combination with other features, represents an important input feature in the automatic classification of topics.¹⁴

Concerning focus, Meurers et al. (2011) report on a pilot study for automatic focus detection. Their goal is Content Assessment, i.e. to assess whether a student’s reply to a text comprehension question answers the question satisfactorily. To this end, they first try to identify the new information in the answer. They suggest that a sentence’s new information focus domain can be determined by first assuming all-new focus (i.e. the new information focus extends over the whole sentence), and then successively excluding given information from the domain.¹⁵ This simple model fails in the case of answers to

¹³The terms sentence and text refer to written language; regarding speech, the corresponding segments are utterance and discourse.

¹⁴Postolache (2005) and Postolache et al. (2005) report on attempts at the classification of sentence topics for Czech. Note, however, that these works are based on the topic definition of the Prague school: *topic* is defined as a non-contrastive contextually bound dependency node, *focus* as a contextually non-bound dependency node, and *contrast* as a contrastive contextually bound dependency node.

¹⁵This procedure needs a definition of givenness extending not only to nominal elements, but also to verbs, adverbs etc. See discussion in Section 2.4.

alternative questions, where the focus consists entirely of given information (see their Example (5)). Yet, it was not conceived as a standalone strategy for focus detection, but to work in combination with other factors, like in sentence topic identification.

- (5) Question: Is the flat in a new building or in an old building?
 Answer: The flat is in a new building.

Another aspect related to discourse givenness is textual cohesion.¹⁶ In Rhetorical Structure Theory (RST, Mann and Thompson (1987), Mann et al. (1993)), tree representations of texts are constructed based on segments of texts (sentences or subtrees) and relations between those segments. An analysis according to RST foresees directed relations from the so-called nucleus (which contains the central information) to the so-called satellite. Directed relations include *elaboration* (the nucleus segment provides basic information, the satellite additional information), *background* (the nucleus is facilitated by the information provided in the satellite, e.g. headings prepare a news article), *preparation* (the satellite is of help to the understanding of the nucleus, e.g. headings prepare for news texts), and many more. Undirected relations, connecting segments on equal terms, are also provided. These so-called *multinuclear relations* include e.g. the *contrast* relation (holding between elements of a juxtaposition). See Figure 1.2 for an example tree structure.¹⁷

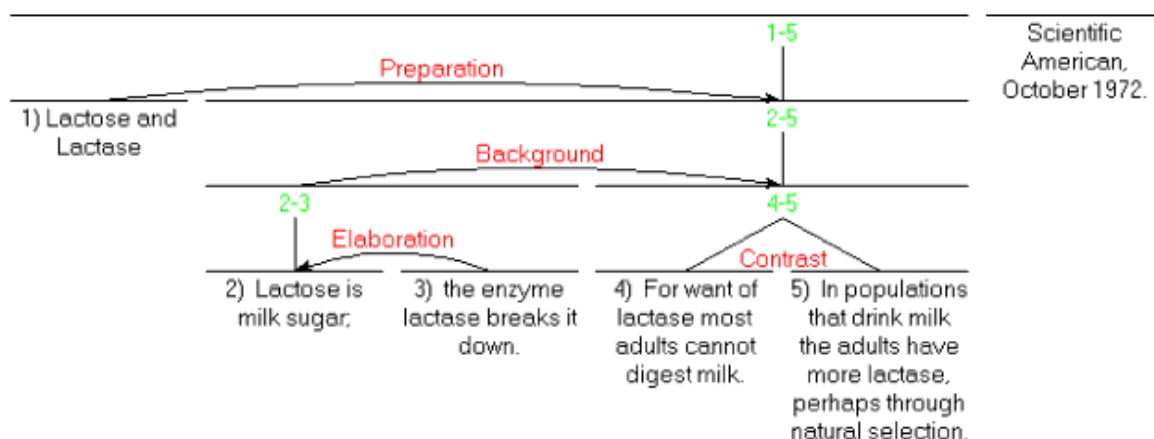


Figure 1.2: Example Structure according to Rhetorical Structure Theory (RST; taken from <http://www.sfu.ca/rst/01intro/intro.html>, last access 27.03.2013).

Reitter (2003), for instance, uses pronouns and lexical similarity as features for the recognition of rhetorical relations. Louis et al. (2010), use features like coreference, givenness, syntactic form and the grammatical role of entities to predict the implicit discourse relation between adjacent sentences.

Research in the area of information structure and rhetorical structure recognition is in a very early stage, but outcomes can be expected to have decisive impact e.g. on Question Answering and Summarization.

¹⁶Asher and Lascarides (2003) record that “rhetorical relations play an essential role in predicting anaphoric bindings” (p. 6).

¹⁷In the graph, the arrows point from satellites to nuclei, allowing to read relations as ‘*x* is a *preparation*/etc. of *y*’.

1.4 The Field of Discourse-Givenness Classification

As mentioned in Section 1.3, the task of discourse-givenness classification has emerged from coreference resolution. Resolving coreference needs $\frac{1}{2}n(n-1)$ comparisons (where n is the number of NPs in the text)¹⁸: every NP_{*i*} needs to be compared to each NP_{*j*} in its preceding context (i.e. position $j < i$). From a very early stage, attempts at limiting the complexity have been made: the number of NPs n can be reduced if one excludes, e.g. NPs that do not refer (e.g. in phrases like *to be on the books*), singletons (referents that are mentioned just once in a text), and discourse-new NPs (which do not have to be compared to NPs to their left; they only need to be taken into account as potential antecedents for NPs to their right).

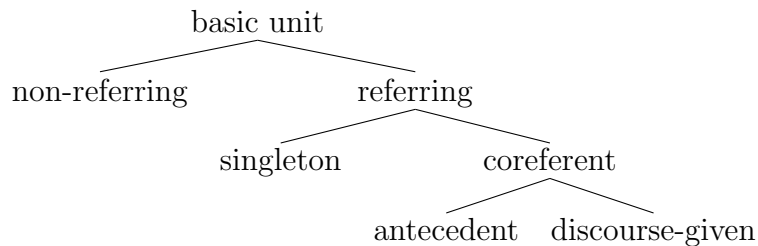


Figure 1.3: Coreference Resolution: Terminology

Thus, coreference resolution can be broken down into a filtering step and a resolution step. The filtering step can be realized with rule-based or learning-based methods; the same holds for the resolution step.

Different divisions of the task into subtasks have been proposed, as shown below: The different approaches are labeled A and B; the subtask of filtering is labeled 1., and the resolution step is labeled 2.

- A.
 1. identification of discourse-given entities
 2. antecedent detection
- B.
 1. identification of referring or coreferent entities (“mention detection”)
 2. building of equivalence classes¹⁹

Variant B has been realized in many contributions to the CoNLL shared task 2011²⁰. However, this variant has a disadvantage: mentions are usually assumed to be noun phrases. This strategy ignores antecedents which are, e.g. whole sentences (see Example (6)). Webber (1988) gives the following example:

- (6) It’s always been presumed that when the glaciers receded, the area got very hot. The Folsom men couldn’t adapt, and they died out₁. That₁’s what is supposed to have happened. It’s the textbook dogma. But it’s wrong. They were human and smart. They adapted their weapons and culture, and they survived.

¹⁸It is usually NPs that are taken as basic units for coreference resolution. This topic is discussed in detail in Section 2.2.1.

¹⁹If the filtering step does not exclude singletons from the set of entities to be resolved, equivalence classes with one element may occur.

²⁰<http://conll.cemantix.org/2011/>, last access March 27th, 2013.

In variant A, in contrast, step 2 allows to add phrases to those identified in step 1. These additional phrases need not meet the same conditions as the phrases in step 1 (e.g. the condition of being an NP).

Approaches to variant B are not considered in this work.

Early approaches to A used a heuristics-based identification of given entities, starting with pronouns (Winograd (1972), Wilks (1973), Hobbs (1976); see Mitkov (1999) for an overview). Refinements to these heuristics, e.g. for excluding pleonastic *it* have been proposed (Paice and Husk, 1987; Kennedy and Boguraev, 1996).

Approaches following these used a wider range of expressions: Hobbs (1993) considers “pronouns, definite noun phrases, and ‘one’ anaphora” (p. 91). Kim and Evens (1996) specialize on proper names. Vieira and Poesio (2000) use syntactically definite descriptions, excluding non-anaphoric definite descriptions on the basis of syntactic and lexical features of the noun phrase. Byron and Gegg-Harrison (2004) take all noun phrases into account, excluding non-anaphoric noun phrases based on syntactic features.

Later, the task of creating a filter was automated: Bergsma et al. (2008), for instance, use distributional methods to exclude non-anaphoric *it*. Evans (2001), Ng and Cardie (2002), Müller (2006) Denis and Baldrige (2007) and Versley et al. (2008a) apply machine learning techniques to the task.

Machine learning techniques for excluding non-anaphoric noun phrases in general have been applied by Ng and Cardie (2002), Ng (2009), Uryupina (2003; 2009) and Zhou and Kong (2011).

Approaches using machine learning methods for more than two categories include Hempelmann’s (2005) model of three-way distinction between *given*, *new*, and *inferable* NPs and Nissim’s (2006) classification of *old* vs. *mediated* vs. *new* NPs (the same task was taken up by Rahman and Ng (2011) and Markert et al. (2012)) and Cahill and Riestler’s (2012) classification of a finer-grained categorization, including experiments with categories conflated to *old*, *mediated*, *new* and *other*.

Research in the field of classification of discourse-givenness focuses mainly on English. Cahill and Riestler’s work on German data forms an exception.

A comparison of existing models for the classification of discourse-givenness and the data they are based on is a desideratum. This will be done in the following chapters.

1.5 Structure of this Document

In Chapter 2, attempts at the formal definition of discourse-givenness and related concepts are presented. The list of controversial phenomena resulting from Chapter 2 forms the basis on which the annotation schemes of existing corpora will be compared and discussed in Chapter 3. These corpora are used to train classifiers for the discourse-givenness of noun phrases. After a review of previous work in the field of discourse-givenness classification (Chapter 4), my own procedure, experimental setup and results are described in Chapter 5. Conclusions and outlooks are provided in Chapter 6.

Chapter 2

Concepts and Definitions

This chapter serves the purpose of a formal definition of discourse-givenness and related concepts (coreference, information status). Previous works in the field of discourse-givenness use definitions that differ from each other, without setting them into relation. The definitions formulated here form the basis of a comparison of annotation schemes which have been used for the classification of discourse-givenness or information status. The structure of this chapter is as follows: first, relevant conceptual and representational conventions of discourse description in general are introduced (Section 2.1). Secondly, a definition of discourse-givenness is provided. This definition is based on the definition of coreference, which in turn is based on the definition of reference. Each concept ‘inherits’ the definitional difficulties and controversial issues of the concept it is based on. For this reason, they will be presented in the following order: definition of reference (Section 2.2), coreference (Section 2.3) and discourse-givenness (Section 2.4). Within the respective sections, attempts at the definition of each concept will be presented, including the respective difficult or controversial issues. These issues will be illustrated with examples (all examples are taken from OntoNotes 1.0 unless specified otherwise). How prevalent these issues are, and how they are resolved in different corpus resources, is addressed in Chapter 3.

A summary is provided at the end of each definition section, which contains the information necessary for the following chapter.

2.1 Basic Concepts and Representation

Mapping discourses to representations of the information they convey needs some conception of how discourse in general functions; this will be given in Section 2.1.1. Based on this, one way of representing this information will be outlined in Section 2.1.2: Discourse Representation Theory (DRT, Kamp and Reyle (1993)).¹ I will assume such formalized representations of discourse as the target representation in a text understanding component. A formalization provides the background for the definitions in the sections that follow.

¹DRT is only one formalism for representing discourse. Alternative formalisms, for instance Heim’s (1982) file change semantics, would also serve the purpose. The choice of formalism is irrelevant to the argumentation.

2.1.1 Basic Concepts in Discourse

Basic elements that make up discourse are words on the syntactic level and concepts on the semantic level. Words stand for concepts, and concepts are meanings that combine into the total meaning of the discourse.

A concept in language can be described in terms of its extension and intension. A concept's *extension* consists of all the objects in the world it is used to describe. A concept's *intension* consists of all the properties that describe the concept, and make an object belong to the set that the concept describes (Washburn, 1898). The prototypical example for concepts with the same intension is *bachelor* and *unmarried man*.² Different intensions can have the same extension. Quine (1986) gives the example of 'cordate' (creature with a heart) and 'renate' (creature with a kidney); both concepts have the same extension (they denote the same set of creatures in the real world), though they are not synonymous.³ The same intension can have different extensions, e.g. at different times: for instance, the concept 'Berliner' (inhabitant of Berlin) in present day describes a different set of people from 'Berliner' say, in 1900. Intensions can be combined to form expressions, which in turn can be used to *refer* to objects in the real world. According to Bach (1987), "to refer to something is simply to express an attitude about it" (p. 52), referring is "part [...] of performing a larger, illocutionary act" (p. 51).⁴ Expressions that refer will be called *referring expressions*, and the objects they refer to in the world will be called *referents*. The meaning or content of a declarative sentence will be called *proposition*.

By means of a discourse, an exchange of information takes place between speaker(s) and hearer(s).⁵ Aiming at a better description of this interactive process, the concept of '*Common Ground*' was conceived by Stalnaker (1974) and Karttunen (1974). Its purpose is "to model the information that is mutually known to be shared and continuously modified in communication" (Krifka, 2008, p. 15). The Common Ground (CG), according to Krifka (2008, p. 16), consists of

- "a set of propositions that is presumed to be mutually accepted (or the conjunction of this set, one proposition)" and
- "a set of entities that had been introduced into the CG before".

Regarding the latter, Krifka elaborates that "[s]uch entities can be explicitly introduced, e.g. by an indefinite NP, or [...] accommodated" (p. 16), and illustrates this with the examples given in (7): (7a) first presents the existence of a cat in the speaker's possession,

²It has been pointed out that there are contexts in which these concepts are not interchangeable (Tye, 1991, p. 144f.), e.g.:

- (1) The pope is an unmarried man.
- (2) #The pope is a bachelor.

Stefanie Dipper (personal communication) suggested using the German words *Samstag* and *Sonnabend* ('Saturday' and 'eve of Sunday') as examples.

³The terms *renate* and *cordate* are not part of any biological classification system, and would not serve the purpose of a taxonomy as they do not discriminate the objects in our world. Soames (2003) notes that Quine introduced these concepts for the purpose of illustration.

⁴A more formal definition of reference will be given in Section 2.2.

⁵Author and reader (for written text) are included, respectively.

and then states that the speaker had to bring this cat to the vet – this order optimizes comprehensibility of the sentence. In (7b), after the utterance of ‘*I had to bring my cat to the vet*’, the recipient accomodates that the speaker owns a cat. Then, that same piece of information is presented explicitly. In this case, the information is redundant because it is already *given*. *Givenness*, according to Krifka, is the category “indicat[ing] that the denotation of an expression is present in the immediate CG content” (p. 37).

- (7) a. I have a cat, and I had to bring my cat to the vet.
 b. #I had to bring my cat to the vet, and I have a cat.

On the basis of “ample evidence that human languages have devices with which speakers can make addressees aware that something that is present in the immediate linguistic context is taken up again” (p. 25), Krifka argues for a *givenness feature*:

- (8) DEFINITION: givenness feature
 “A feature X of an expression α is a Givenness feature if X indicates whether the denotation of α is present in the CG or not, and/or indicates the degree to which it is present in the immediate CG.” (p. 25)

Regarding anaphoric expressions in particular, he describes them as “specific linguistic forms that indicate the givenness status of their denotations, including personal pronouns, clitics and person inflection, demonstratives, definite articles [...]. Definite articles can be used to indicate whether a denotation is given in a CG in general, whereas clitics and pronouns typically indicate that their denotations are given in the immediate CG [and] indefinite articles [...] indicate that their referent is not given” (p. 27).

For the classification task, this observation suggests that the expression’s form can provide important clues for determining its discourse-givenness or information status.

2.1.2 Formal Representation

Discourse updates and the truth conditions of a discourse can be described by representation structures. In DRT (Kamp and Reyle, 1993), such structures are called Discourse Representation Structure (DRS). For example DRSs, see Figures 2.1 to 2.3; these DRSs are adapted from (Kamp and Reyle, 1993, pp. 69, 310, and 332) and will be explained in the following.

DRSs consist of

- (i) a set of *discourse referents* (notated in the separated top parts of the boxes) and
- (ii) a set of *conditions* representing the information conveyed by the discourse (notated in the main parts of the boxes).

Discourse referents (individuals/sets) are an intermediate representation between referring expressions and referents in the world. They are notated as variables.⁶ *Conditions* are formulated by combining these variables with relations. These relations represent

- (i) properties of individuals (or sets) and relations in the real world, such as the predicate *book* in Figure 2.2 or the relation *owns* in Figure 2.1, or

⁶In some of the literature, proper names are represented by constants rather than variables.

- (ii) logical operators (e.g. junctors like *and*, *if-then* ‘ \Rightarrow ’) or
- (iii) quantifiers: quantifiers express either the existence of an object or proportional relations in the real world, e.g. the universal quantifier ‘ \forall ’ or the quantifier ‘most’ as in Figure 2.3.

Conditions are implicitly combined using the logical operator *and*.

<i>x y u v</i>
Jones(x)
Ulysses(y)
x owns y
u = y
v = x
u fascinates v

Figure 2.1: DRS of ‘Jones owns Ulysses. It fascinates him.’ (taken from Kamp and Reyle (1993), p. 69)

A *mention* (usually a noun phrase) introduces a variable into a discourse representation structure. There are different kinds of variables:

- (i) those for atomic discourse referents (representing individual objects), which are represented by lower case letters (*x*, *y*, *u*, and *v* in Figure 2.1),
- (ii) those for non-atomic discourse referents (representing sets of individuals), which are represented by upper case letters (such as *Y* and *U* in Figure 2.2), and
- (iii) those for vague discourse referents (neither atomic nor non-atomic, e.g. bare plurals inside the scope of a quantifier), which are represented by lower case Greek letters (e.g. η in Figure 2.3⁷).

Variables for semantically definite descriptions are always introduced at the top level (the main DRS), see Figures 2.1 (variables *x*, *y*, *u*, *v*) and 2.2 (variables *x*, *z*, *U*, *t*, *w*). Variables for indefinite descriptions are introduced at the nesting level they are processed (variable η in Figure 2.3), and universally quantified NPs (NPs beginning with ‘every’) are introduced at a level subordinate to that at which they are processed (variables *y* in Figure 2.2 and *x* in Figure 2.3).

In the case of coreference, the newly introduced variable is linked to the context by the identity relation (e.g. ‘ $x = y$ ’), see the bold faced conditions in Figures 2.1, 2.2 and 2.3. For pronouns, Kamp and Reyle (1993) provide two alternative suggestions: either to introduce a discourse referent and stipulate identity (‘ $x = y$ ’), or to replace the pronoun by the referent variable that is chosen as an antecedent.

The validity and appropriateness of these DRSs is checked against models (each model instantiating a ‘possible world’). A model “is a certain information structure, relative to

⁷The predicate *book* is marked with a star (‘*’) here. The star notates that variant of a predicate which may also be applied to a set of individuals (*book(X)*), instead of one individual at a time (*book(x)*).

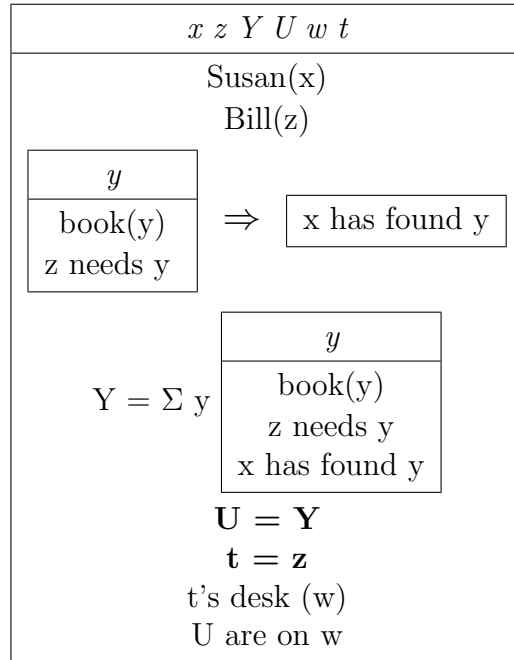


Figure 2.2: DRS of ‘Susan has found every book which Bill needs. They are on his desk.’ (taken from Kamp and Reyle (1993), p. 310)

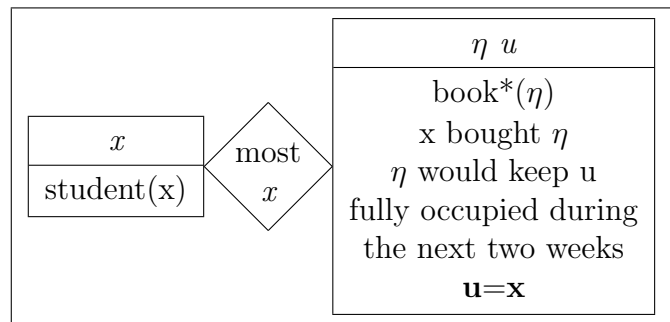


Figure 2.3: DRS of ‘Most students bought books that would keep them fully occupied during the next two weeks.’ (taken from Kamp and Reyle (1993), p. 332)

which it is possible to evaluate the expressions of some given language, and in particular to evaluate the sentences of that language in respect of truth and falsity” (Kamp and Reyle, 1993, p. 93). Models consist of objects and relations between them (properties can be regarded as unary relations).

For the evaluation, each discourse referent in a DRS K is mapped to an element in U_K , the universe of K . A DRS “ K is true in [model] M iff there are corresponding to the members of u_1, \dots, u_n of U_K objects a_1, \dots, a_n in M such that the conditions in K are satisfied in M by the objects which correspond to the referents that these conditions contain as arguments” (Kamp and Reyle, 1993, p. 130).

The DRS in Figure 2.2, for instance, would be true evaluated with respect to model M_1 ⁸:

- (9) M_1 :
 $U_{M_1} = \{a, b, c, d, e, f\}$
 $Name_{M_1} = \{\langle Susan, a \rangle, \langle Bill, b \rangle\}$
 $Pred_{M_1}$:
 $Pred_{M_1}(book) = \{c, d, e\}$;
 $Pred_{M_1}(need) = \{\langle b, c \rangle, \langle b, d \rangle\}$;
 $Pred_{M_1}(has\ found) = \{\langle s, c \rangle, \langle s, d \rangle, \langle b, e \rangle\}$;
 $Pred_{M_1}(x's\ desk) = \{\langle b, f \rangle\}$;
 $Pred_{M_1}(be\ on) = \{\langle c, f \rangle, \langle d, f \rangle\}$

$Susan(x)$, for instance, evaluated with respect to this model, is mapped to a :
 $\llbracket Susan(x) \rrbracket_{M_1, t} = a$.

The formalization presented in this section will be used throughout the theoretical part of this work. It enables a precise definition of coreference; a formal representation of discourse provides a basis for its further computational processing.

2.2 Reference

An expression is termed ‘referring’ if it designates (i.e. stands for) some entity in the real world: a concrete object, state, event, etc. or set thereof, cf. Clément (2000), Bußmann (2000).⁹ This entity (or set of entities, respectively) is called the referent. Bach’s (1987) definition is given in (10).¹⁰

- (10) DEFINITION
 “To refer to something is to use a singular term with the intention (part of one’s communicative intention) of indicating to one’s audience the object of the attitude one is expressing” (p. 52). A singular term is “any expression that can be used to refer to an individual, whether or not it [determinately] denotes” (p. 62).

Bach suggests to represent reference as “a four-place relation between a speaker, a word [or other expression], an audience and an object” (p. 39; see (11) for a formalized version).

- (11) DEFINITION
 $Ref = \{\langle s, e, a, o \rangle \mid \text{speaker } s \text{ uses expression } e \text{ in front of audience } a \text{ to refer to object } o\}$

He acknowledges suggestions of alternative definitions¹¹, which include the expression’s context, e.g. one which could be formalized as shown in (12). These definitions he rejects, arguing that reference is not a fully determinate function, even if textual and situative context (such as discourse participants, location, etc.) were included in the parameter c .¹²

⁸Abbreviations: U - universe, M - model, Pred - Predicate(s).

⁹Entities in fictional worlds (e.g. book or film characters) will also be assumed referents here. The issue of the world of reference is discussed in Section 2.2.3.

¹⁰It can be argued that this definition is redundant, if not circular.

¹¹Unfortunately, he does not provide references for these definitions.

¹²In my opinion, the argument is invalid: if reference is not a function, and be partially changed into a function by adding a certain parameter, it does not follow that this additional parameter is useless in general. It is true however that neither definition results in a function.

However, he points out that the context plays a role in the interpretation of expressions that possibly refer: “referring never occurs by itself (...) [but] is always part and parcel of performing a larger, illocutionary act” (Bach, 1987, p. 51).

(12) DEFINITION

Ref= $\{ \langle e, c, o \rangle \mid \text{expression } e \text{ in context } c \text{ refers to object } o \}$

Van Deemter and Kibble (2000) use Bach’s definition (11) in a functional notation (see (13)) as a means to later define coreference. This notation “suppresses the role of context”, assuming that the expressions involved “have a unique referent in the context in which they occur (i.e., their context in the corpus makes them unambiguous)” (van Deemter and Kibble, 2000, p. 629). Problematic aspects of this notational trick manifest themselves when this definition is used in the definition of coreference; they are discussed in Section 2.3. Expressed formally, for getting a functional relation, the function’s domain needs to be restricted to unambiguous specific expressions α .

(13) DEFINITION

Ref(α)= $\{ r \mid r \text{ is the entity referred to by } \alpha \}$

In my opinion, reference is a relation as defined in (14) – an extension of (12) –, where the context c extends to both textual and situative context, including speaker and audience. Neither of these relations, however, is a function.

(14) DEFINITION

Ref= $\{ \langle e, c, r \rangle \mid \text{expression } e \text{ in context } c \text{ refers to an object } r \text{ with } r \in D(e), \text{ or a set of objects } r \text{ with } r \subseteq D(e), \text{ where } D \text{ is the denotation of } e \text{ (excluding determiners).} \}$

A tentative distinction of types of reference is given in Table 2.1; some cases of reference may show characteristics of more than one type. A stricter definition of reference, specific reference, which includes subtypes 1, 2 and 4, is discussed in Section 2.2.4.

-
1. $D(e) := r$ (**e is used for r by convention**)
names (unique, e.g. *John Gill* or less unique, e.g. *John*) and nicknames (including uses of expressions where the literal meaning of the signifier does not hold, e.g. using *the professor* for referring to the musician Benny Goodman)
 2. $D(e) = r$; **e is a description sufficient for r ’s unique identification**
e.g. *the first man on the moon* and deictic expressions ($D_c(e) = r$ in context c)
 3. **$the\ set\ r = D(e)$; e is used to distinguish $x \in D(e)$ from $y \notin D(e)$**
generalizations (e.g. *Cats need calcium*) are discussed in Section 2.2.5
 4. $r \in D(e)$ or **$set\ r \subset D(e)$, disregarding other $x \in D(e)$ and $y \notin D(e)$**
types (2.) and (4.) have some overlap, as referents of expressions like *the dean* in a conversation between two students are uniquely identifiable with little background knowledge about that part of the world the discourse is about
-

Table 2.1: Types of Reference (Tentative)

On a final note, referents may remain ‘anonymous’, i.e. the hearer might not identify the referent(s) in the world: Allan (2012) notes that for an expression to be used referringly, “[p]hysical identification is not necessary, a hearer only needs to have a cogent grasp of what differentiates the speaker’s (presumed) referent from any distractors” (p. 20; cf. also Strawson (1950) and Donnellan (1966)).

2.2.1 Identifying Basic Units

In this section, the question of what constitutes a referring expression will be addressed. Typically, referring is associated with (but not limited to) noun phrases, including names and pronouns. Karttunen points out a “suggestion by Noam Chomsky (1965) [...] that the base component of a transformational grammar associates with each noun phrase a referential index, say, some integer” (Karttunen (1969a, p. 5), see (15)¹³)¹⁴.

- (15) TASK DEFINITION
Associate a referential index (e.g. an integer) with each noun phrase. (Chomsky, 1965; Karttunen, 1969a)

Consider the text given in (16), analyzed according to this suggestion.¹⁵

- (16) Four former Cordis Corp.₁ officials₂ were acquitted of federal charges related to the Miami-based company’s₁ sale of pacemakers_{3,4,5}, including conspiracy to hide pacemaker defects_{6,7}. Jurors in U.S. District Court₈ in Miami_{9,8,10} cleared Harold Hershenson₁₁, a former executive vice president_{att:11,11}; John Pagonis₁₂, a former vice president_{att:12,12}; and Stephen Vadas₁₃ and Dean Ciporkin_{14,15}, who₁₅ had been engineers with Cordis_{1att:15,15,2}.

From this analysis, we observe that referring expressions can occur embedded into other referring expressions, e.g. *Cordis Corp.* in *Four former Cordis Corp. officials*. We also observe expressions which embed expressions with the same referential index, e.g. indices 8 (*U.S. District Court in Miami* containing *U.S. District Court*), 11 (*Harold Hershenson, a former executive vice president* containing *Harold Hershenson*) and 15 (*Stephen Vadas and Dean Ciporkin, who had been engineers with Cordis* containing *Stephen Vadas and Dean Ciporkin* and *who*). According to this analysis, one constituent would contain two or more mentions of the same referent. Moreover, the relative pronoun (index 15) only serves the purpose of linking between two parts of the same constituent. Categorizing it

¹³Chomsky supposes that “certain lexical items are designated as ‘referential’ and that by a general convention each occurrence of a referential item is assigned a marker, say, an integer, as a feature” (Chomsky, 1965, p. 145).

¹⁴There are two equivalent options of handling indices: (1) introduce a new integer for each expression, later replace those referring to the same referent by the index of the first mention or (2) re-use an integer for all expressions referring to the same referent. Option (2) is applied here for reasons of representation.

¹⁵For the sample text in (16), an intuitive understanding of referring vs. non-referring expressions and of coreference is sufficient. Definitions are given in Sections 2.2.2 (non-referring expressions) and 2.3 (coreference). Attributions, which are non-referring in general, are marked with *att* and the index of the expression they apply to. Noun phrases are annotated to include possessive ‘s’ if applicable (*the Miami-based company’s*).

as a *mention* of the referent set {Stephen Vadas, Dean Ciporkin} does not seem adequate. Thus, in the case of embedding with identical indices, I argue for the annotation of the maximal referring expression only.

Another question is whether or not it is necessary for a mention of a referent to form a noun phrase in its own right: the referent Miami (indexed 9) – the venue of the legal dispute – has been mentioned earlier as a part of the word *Miami-based*. Similarly, the word *pacemaker* as a premodifier of *pacemaker defects* could be interpreted as referring to a set of pacemakers (either the same set as, a subset of, or a set intersecting with the pacemakers indexed 3): the compound *pacemaker defects* can be rephrased as either *defects in pacemakers* or *defects in the pacemakers*. The text implies the existence of a set of pacemakers, which were meant to be sold – with the fact that they were defective concealed or unknown.

Malte Zimmermann (personal communication) suggests the following referentiality test:

(17) A noun in a compound is referring if and only if it can be taken up with a pronoun.

See his minimal pair example in (18). Referentiality of nominal modifiers (indexed 1 in Examples (18) and (19)¹⁶) may be dependent on their compositional relation to the noun they modify (indexed 2 in the examples): it seems that a noun n_1 can be taken up more easily if it occurs as a *subject* premodifier of a deverbal noun n_2 than if it was an *object* premodifier of noun n_2 . The exact conditions need further investigation. This, however, is outside the scope of this work. It is only important here to state as a result that it is possible for premodifiers to refer.

(18) a. # Lion_{1^{sg;obj}} hunting₂ is fun because they_{1^{pl;subj}} are dangerous animals.

b. Hunting₂ lions_{1^{pl;obj}} is fun because they_{1^{pl;subj}} are dangerous animals.

(19) a. Sheep_{1^{pl;subj}} grazing₂ is unsuitable because they_{1^{pl;subj}} selectively feed on the foodplant and can reduce or even eliminate it from sites.

b. The distances involved in bird_{1^{sg;subj}} migration₂ mean that they_{1^{pl;subj}} often cross political boundaries of countries and conservation measures require international cooperation.

In principle, referents do not have to be explicitly mentioned, e.g. in pro-drop languages (see (20) for a Spanish example; the dropped pronouns are marked with subscript *dropped* in the English translation below).¹⁷ This, however, is beyond the scope of this work.

(20) *Pedro trabaja en una mina de oro. Un día será rico, pero no lo sabe aún.*
 Pedro work.3pers in a goldmine. One day, be.3pers.fut rich, but not it know.3pers yet.

Pedro works in a goldmine. One day, he_{dropped} will be rich, but he_{dropped} does not know it yet.

¹⁶Sources: http://butterfly-conservation.org/files/bcw_narrow-bordered-bee-hawk-moth-nbbh_eng.pdf and http://en.wikipedia.org/wiki/Bird_migration (last access April 9th, 2013). Sentence 19b occurs after ‘*Human activities have threatened many migratory bird species.*’ An interpretation of *they* in (19b) referring back to *many migratory bird species* is implausible (it is not species who cross political boundaries but groups of objects).

¹⁷Abbreviations in glosses: 3pers - 3rd person; fut - future tense.

Referents can be mentioned in the form of sentences and even larger sections of text (cf. (Webber, 1988; Dipper and Zinsmeister, 2010)): Propositions and sets of propositions can be referred back to like in Examples (21) (taken from Dipper and Zinsmeister (2010), p. 56) and (22) (taken from the Trains-91 section of the ARRAU corpus).

- (21) I would like to draw particular attention to the fact that people who have made their lives here in the European Union_{1;part1} still do not have the right to vote_{1;part2}, even though the European Parliament has called for it₁ on many occasions.
- (22) S : okay ss okay so we have all right let 's see the E1 goes to Bath and picks up a boxcar and comes back then loads bananas and goes to Corning via Dansville and drops off the bananas ' n meanwhile engine 2 takes a boxcar to Corning_{1;part 1}
M : m hm
S : loads the oranges hooks up the tanker goes to Elmira makes the OJ and then goes back to Avon via Bath_{1;part 2}
M : right
S : okay now that₁ would work except we just found out that the boxcar at Elmira is n't working and will n't be repaired until 8 AM

There is consensus that reference of such complex forms is different from reference of NPs (cf. Webber (1988), Asher (1993), Byron (2002) and Schwarz-Friese et al. (2007), among others). Complexly-formed referents are introduced into the discourse representation only if they are taken up again.

I suggest a reformulation of the task, based on Chomsky's suggestion (see (15)), as follows:

(23) TASK DEFINITION

Consider each noun phrase and compositional part of compound for the following steps. Discard nonreferring uses of expressions. Enumerate and provide with an index

- (i) all different referents, i.e. entities mentioned in the text (if not sure about identity of referent, create new index), and
- (ii) all different referring expressions, i.e. all mentions of each referent (if one mention embeds a mention of the same referent, only use the maximal referring expression).

Which uses of expressions are considered nonreferring is discussed in the following section. Performing step (a) on example text (16), we get the following list of referents:

1. Cordis Corp., a seller of pacemakers, based in Miami
2. Harold Hershenson, a former executive vice president of 1.
3. John Pagonis, a former vice president of 1.
4. Stephen Vadas
5. Dean Ciporkin

6. {4., 5.}, former engineers with 1.
7. {2., 3., 4., 5.}, four former officials of 1.
8. pacemakers
9. the sale of 8. by 1.
10. pacemakers (probably a subset of or a set intersecting with 8.)
11. defects in 10.
12. U.S. District Court in Miami
13. Miami, location of 12, base of 1.
14. Jurors in 12.
15. charges of 7. by 14., related to 9.
16. alleged conspiracy of 7. to hide 11. (16. is subset of 15.)

Performing step (ii) results in the following list of referring expressions (the first number represents the referent's index, the second number represents the mention's index):

- 1.1. Cordis Corp.
- 1.2. the Miami-based company's
- 1.3. Cordis
- 2.1. Harold Hershhenson, a former executive vice president
- 3.1. John Pagones, a former vice president
- 4.1. Stephen Vadas
- 5.1. Dean Ciporkin
- 6.1. Stephen Vadas and Dean Ciporkin, who had been engineers with Cordis
- 7.1. Four former Cordis Corp. officials
- 7.2. Harold Hershhenson, a former executive vice president; John Pagones, a former vice president; and Stephen Vadas and Dean Ciporkin, who had been engineers with Cordis
- 8.1. pacemakers
- 9.1. the Miami-based company's sale of pacemakers
- 10.1. pacemaker (in the NP *pacemaker defects*)
- 11.1. pacemaker defects

12.1. U.S. District Court in Miami

13.1. Miami (in the adjective *Miami-based*)

13.2. Miami (in the NP *Jurors in U.S. District Court in Miami*)

14.1. Jurors in U.S. District Court in Miami

15.1. federal charges related to the Miami-based company's sale of pacemakers

16.1. conspiracy to hide pacemaker defects

Performing the steps in this order helps to identify expressions that are less obviously referring (like the first mention of *Miami* in Example (16) and the sentence(s) in Examples (21) and (22)).

2.2.2 Non-Referring Use of Expressions

Expressions used non-referringly include expletive pronouns (Example (24a)) and expressions in phrases with non-compositional meaning, such as collocational expressions (Examples (24b) and (24c)) and idiomatic expressions (Examples (24d) to (24g)): there is no concrete object, role, etc. involved in the utterance situation of these sentences (NPs under discussion are underlined and marked *nonref*).

- (24) a. It_{nonref} seems to me that this is the pin that has finally pricked the balloon [...]
- b. [...] the ministry played a role_{nonref} in orchestrating recent moves by Japanese banks
- c. Certainly the federal government should take a hard look_{nonref} at it.
- d. It should be noted that Mr. Schwartz (...) is a puckish sort who likes to give his colleagues the needle_{nonref}.
- e. Bank officials, however, showed him the door_{nonref}, and the sale never came off.
- f. [...] line supervisors slice up the merit pie_{nonref}.
- g. Gives me the willies_{nonref} just thinking about it.

Expressions in their attributive use also do not refer (Donnellan, 1966).¹⁸ Explicit uses of attributions are predications, appositions, and attributive relative clauses, see Examples (25a) to (25c) (marked *att*) as well as the second sentence in Example (16).¹⁹

- (25) a. The ULI is a non-profit research and education group based in Washington, D.C. [...] _{att}.

¹⁸Kamp and Reyle (1993) treat attributives with a definite determiner as referring, see Section 2.3.4.

¹⁹In OntoNotes, adjective phrases can also form appositions, e.g. *46 years old* in Example (25b) and *Born in a Baltic town...* the following example:

- (1) Born in a Baltic town in an area which is now part of Poland_{att}, he has dedicated his life to the party apparatus.

- b. Mr. Amon, 46 years old_{att}, is the company's director of quality assurance _{att} [...].
- c. Jurors in U.S. District Court in Miami cleared [...] Stephen Vadas and Dean Ciporkin, who had been engineers with Cordis_{att}.

Donnellan (1966) points out that definite descriptions²⁰ like *Smith's murderer* (see his example in (26)) can be used referringly or attributively (the latter meaning ‘whoever has intentionally killed Smith’).²¹ These cases, however, are extremely hard to distinguish: definite descriptions do not always give away the referent’s identity; their referents may or may not be uniquely identifiable. Thus, to distinguish whether the speaker has a certain referent in mind, exact information on the speaker’s model of the world is needed. For the annotation process, the referring use can be assumed as the default case. The attributive use is only annotated if the context makes it explicit.

(26) Smith's murderer_{referring/attributive} is insane.

2.2.3 World of Reference

Reference is evaluated with respect to a certain world. First attempts at a definition of reference used existence in the real world as a criterion for reference. However, it is also possible to make propositions about objects that do not exist in the real world, but in a hypothetical or fictional world, see Examples (27) and (28).

(27) If Lucy had a cat, she'd name it Pebbles.

(28) Watson is Sherlock Holmes's assistant.

The world of reference is not made explicit. Consider Karttunen's example in (29) (Karttunen, 1968, p. 21). Sentence (29a), considered on its own, is ambiguous. If it is continued with sentence (29b), the referent exists in the real world. If continued with (29c), however, it exists only in Mary's imaginative world. In the DRS in Figure 2.4, the variable *y* is part of the main DRS, whereas it is part of the inner DRS in Figure 2.5.

(29) a. Mary wants to marry a rich man₁.

b. He₁ is a banker.

c. He₁ must be a banker. (in the sense of ‘It is necessary that he be rich.’)

Karttunen (1969a, p. 34) postulates that “an indefinite NP establishes a discourse referent just in case the sentence is an affirmative assertion”. He assumes certain “‘world-creating’ verbs” (p. 34), such as *intend* and *want* (also called intensional verbs). In Donnellan's (1966) categorization, (29b) would invoke a referring, (29c) an attributive reading.

Referring expressions in conditional sentences can be interpreted in two different ways in DRT. Consider the example sentences (30) and (31)²² and the corresponding DRSs (2.6

²⁰Definite descriptions are noun phrases of the form ‘the N’ where N is a singular common noun or a noun phrase (cf. Ludlow (2011)). The definite determiner may be replaced by a possessive.

²¹This distinction resembles the distinction between *de dicto* and *de re*: a *de dicto* use is one where an object or set of objects must meet the description; a *de re* use is one where an object or set of objects happen to meet the description.

²²Example (30) is taken from OntoNotes 1.0, (31) is constructed to form a minimal pair.

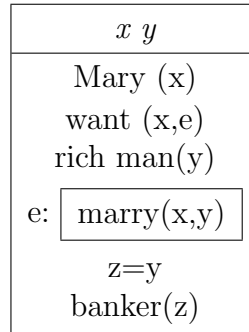


Figure 2.4: DRS of ‘Mary wants to marry a rich man. He is a banker.’

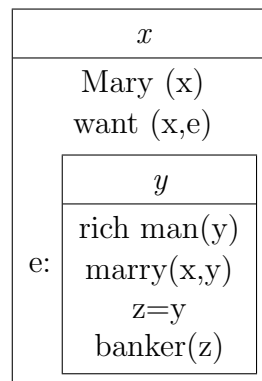


Figure 2.5: DRS of ‘Mary wants to marry a rich man. He must be a banker.’

and 2.7, respectively). Regarding the semantic content, they are equivalent in meaning. Regarding the pragmatic content (i.e. also the presuppositions), they differ: in Figure 2.6, the discourse referent for the set of parents is declared in the inner DRS. In Figure 2.7, it forms part of the main DRS, presupposing the existence of parents.

- (30) If parents₁ are dissatisfied with a school, they₁ should have the option of switching to another.
- (31) Parents₁ should have the option of switching to another school, if they₁ are dissatisfied with the current school.

It is important to note that DRSs are designed to represent the semantic content of discourses. As to presuppositions, the “level of accomodation [...] is also pragmatically determined” (Asher and Lascarides, 1998, p. 287), i.e. discourse is interpreted with the help of world knowledge. In most cases, like ‘parents’ above, it is clear whether a presupposition is plausible or whether it has to be treated with caution. Regarding presuppositions in fictional texts, Asher and Lascarides state that “[t]here is presupposition failure but the reader doesn’t really care” (Asher and Lascarides, 1998, p. 287).²³

²³A predicate’s actual extension or temporal dimension only becomes apparent when the DRS is mapped to a model.

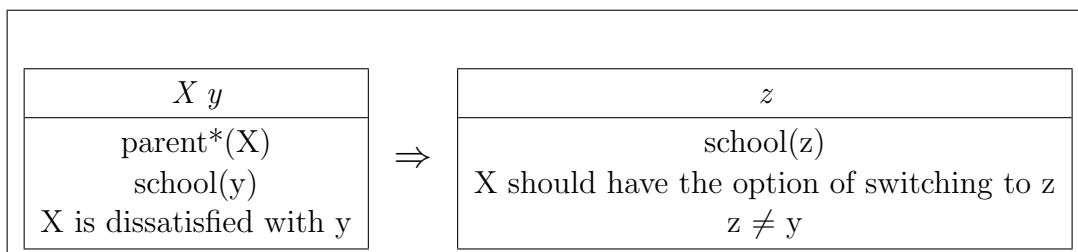


Figure 2.6: DRS of ‘If parents are dissatisfied with a school, they should have the option of switching to another.’

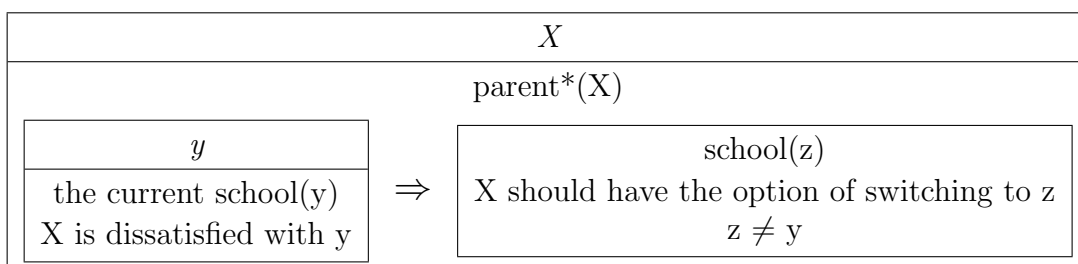


Figure 2.7: DRS of ‘Parents should have the option of switching to another school, if they are dissatisfied with the current school.’

2.2.4 Specificity

In this section, the concept of specificity (semantic definiteness) will be discussed. The purpose of this concept is to narrow down referring expressions to those expressions with a determinate referent, so reference can be considered a function. This function is used for the definition of coreference in Section 2.3.

An expression is specific or semantically definite if it is used to select an individual object or set of objects against other objects of the same kind, to “single out the entity uniquely” (van Deemter and Kibble, 2000, p. 631). Van Deemter and Kibble (2000) elaborate that

- (32) “[a] semantically definite NP α is one whose set-theoretic denotation takes the form of a principal filter (Partee et al., 1990), i.e., a set of the form $X : Y \subseteq X$ for some set of individuals Y ” (p. 635).

Example (33) contains references to four specific bridges or sets of bridges: *the nation’s old bridges* (subscript index 1), *G Street Bridge* (index 2), *Charter Oak Bridge* (index 3), and *a bridge* in Peninsula, Ohio (index 4). The expressions underlined but not indexed, in contrast, are non-specific uses involving the concept of bridges: *Bridges* and *older bridges* refer to bridges (or older bridges, respectively) in general; *just an ugly bridge* and *one that blocks the view of a new park below* are predications applying to *G Street Bridge*. Noun phrases that are subject to generalizations (quantified, negated or kind-referring expressions), as well as vague and ambiguous reference will be discussed in detail in the following sections.

(33) Beauty Takes Backseat To Safety on Bridges

Everyone agrees that most of the nation's old bridges₁ need to be repaired or replaced. But there's disagreement over how to do it. Highway officials insist the ornamental railings on older bridges aren't strong enough to prevent vehicles from crashing through. But other people don't want to lose the bridges' beautiful, sometimes historic, features. "The primary purpose of a railing is to contain a vehicle and not to provide a scenic view," says Jack White, a planner with the Indiana Highway Department. He and others prefer to install railings such as the "type F safety shape," a four-foot-high concrete slab with no openings. In Richmond, Ind., the type F railing is being used to replace arched openings on the G Street Bridge₂. Garret Boone, who teaches art at Earlham College, calls the new structure₂ "just an ugly bridge_{att:2}" and one that blocks the view of a new park below_{att:2}. In Hartford, Conn., the Charter Oak Bridge₃ will soon be replaced, the cast-iron medallions from its₃ railings relegated to a park. Compromises are possible. Citizens in Peninsula, Ohio, upset over changes to a bridge₄, negotiated a deal: The bottom half of the railing will be type F, while the top half will have the old bridge's₄ floral pattern.

An expression is specific even if the actual identity of the referent (its extension) is not revealed, as the NP indexed 4 in Example (33) above, and the NP indexed 1 in Example (34). An indefinite first mention may be non-specific to the hearer (and even the speaker may not know the exact identity of the referent) like in Example (29a). Later mentions of this referent, or an occurrence in an event-presenting sentence (marked by the past tense or the present progressive) like in Example (34), however, may make it clear that the expression is used specifically. In file change semantics²⁴, indefinites systematically introduce new file cards: "For every indefinite, start a new card. For every definite, update an old card" (Heim, 1982, p. 227).

(34) Last year a gunner₁ shot a whooper by mistake thinking that it was a snow goose. He₁ paid an immense fine [...].

Enç (1991) and von Heusinger (2002), among others, have observed that there is neither consensus on the notion of specificity nor on definitional criteria. Terminology is used inconsistently: what von Heusinger calls specificity is termed *referentiality* by Givón (1978). Criteria like identifiability of the referent by the speaker and the existential presupposition have been postulated and later given up (von Heusinger, 2002). Enç (1991) defines specific expressions as expressions "linked to previously established discourse referents" (p. 9). According to his definition, (syntactically) definite expressions are those linked to an antecedent by the *identity* relation. Specific expressions are those linked by the *inclusion* relation (being a subset or standing in "some recoverable relation" (Enç, 1991, p. 24) to the antecedent). Von Heusinger (2002) terms the latter expressions *relative specifics*, and the expressions they are linked to *anchor*. The reference of the expression "depends on the 'anchor' expression [in that] [o]nce the reference for the anchored is determined, the reference for the specific term is also determined" (von Heusinger, 2002, p. 36).

²⁴File change semantics uses the metaphor of file cards: all the information on a referent is stored on the referent's file card.

There are two problems with this definition: first, unanchored specific expressions (e.g. first mentions) are unprovided for. Second, the definition includes anchored non-specific expressions: an expression which is anchored to a non-specific antecedent would have to be categorized as specific, because its reference is identical to or can be determined relative to its antecedent. This is the case in Example (29a) continued with *c* (*Mary wants to marry a rich man. He must be a banker.*), and the examples in (35). I prefer to categorize these examples as non-specific, because their reference is dependent on the reference of the antecedent; the antecedent does not refer to a single concrete object or event.

In particular, these definite expressions (indexed 2 in Example (35a), and 2 and 3, respectively in (35b)), should be categorised as *generic*. This will be explained in the following section.

- (35) a. Running for president in early 1980, he was also quoted as supporting federal funding for abortions₁ in cases of rape, incest and to save the life of the mother₂; *specific relative to 1*.
- b. When husbands₁ take on more housework, they tend to substitute for chores done by the kids₂; *specific relative to 1* rather than by the wife₃; *specific relative to 1*.

Another issue of Enç and von Heusinger’s anchoring-based definition of specificity is that if it is integrated as a precondition into a definition of coreference, the whole definition becomes circular²⁵. For this reason, I give a negative definition of specificity, excluding non-specific expressions.

Non-specific expressions challenge the modeling of reference as a function: a function should return exactly one referent, which can be delimited against other referents in the text.

Non-specific expressions include expressions that are subject to generalizations, as well as vague and ambiguous expressions. These kinds of expressions are discussed in the following.

2.2.5 Generalizations

One type of expressions complementary to specific reference is generalizations, i.e. statements that abstract from concrete events. Generalizations include negation (36a), quantification (36b), and genericity ((36c) to (36f); (36f) from Krifka et al. (1995)).

- (36) a. No lawyers or tape recorders were present.
- b. Every issue is multisided.
- c. [...] trout have very soft mouths.
- d. Underclass neighborhoods offer relatively few employment opportunities [...]
- e. Small neighborhood businesses could provide more jobs, if crime were not so harmful to creating and maintaining those businesses.
- f. The lion has a bushy tail.

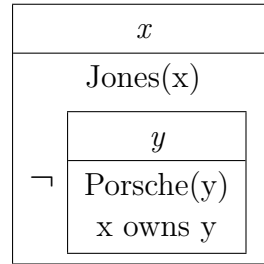


Figure 2.8: DRS of ‘Jones does not own a Porsche.’ (taken from Kamp and Reyle (1993), p. 102)

For negation, consider the DRS in Figure 2.8. Note that the variable for *Porsche* is under the scope of the negation, whereas the variable for *Jones* is not. For quantification, reconsider the DRSs in Figures 2.2 and 2.3: the variables y for *every book* in (2.2) and x for *most students* are under the scope of quantifiers. They are not introduced at the main level but at a subordinate level, and thus they do not form part of the representation of the main world of reference.

Generic expressions can also be analyzed as binding variables, see Figure 2.9 (Kamp and Reyle, 1993, p. 294). The generic operator is represented by a wavy arrow (\approx), similarly to the arrow representation of the universal quantifier.

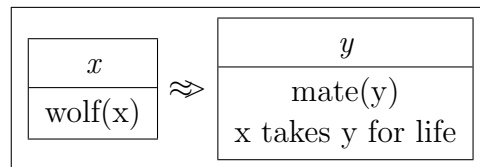


Figure 2.9: DRS of ‘A wolf takes a mate for life.’ (taken from Kamp and Reyle (1993), p. 294)

There is controversy over which variables are bound in generalizations, and also how many variables are bound. Krifka et al. (1995) differentiate

- (a) utterances containing an expression referring to a kind (as opposed to expressions referring specifically, i.e. to concrete objects), see Example (36a), and
- (b) those expressing a “regularity which summarizes groups of particular episodes or facts” (p. 2).

Genericity can be tested in the following way²⁶:

²⁵An expression is coreferent if it is specific and its referent is identical to its antecedent’s referent. An expression is specific if it is coreferent with or related to another expression in the text.

²⁶If the entities to be tested are already of the form suggested in the test, they can be directly assigned the respective class.

- (37) Type (a): Insert ‘as a whole’ or ‘in general’ after the noun phrase you wish to test for genericity. If the sentence maintains its meaning, then the noun phrase refers to a kind. Otherwise apply the following test:
- (38) Type (b): Insert ‘generally’ or ‘typically’ into the sentence. If it maintains its meaning, then it is a regularity (or characterization in Cohen’s (2001) terms).

Another issue, which is discussed controversially, is what exactly the expressions of type (a) denote: the property (intension), the set of individuals (extension), a quantified (sub)set (*most/some* or *all*), etc. For discussions, see Carlson (1977b), Cohen (2001), Cohen and Erteschik-Shir (2002), Krifka (2004) and Herbelot (2011).

Lawler (1972), Dahl (1975), and Carlson (1977a; 1977b), among others, suggest an operator for genericity which binds one variable (a monadic genericity operator), namely y in (39) (adapted from Krifka et al. (1995), p. 20 and 22, marked as a “tentative rule” (Krifka et al., 1995, p. 22)). Heim (1982) and also Carlson in his later work (1989) suggest a dyadic operator, i.e. an operator with two open propositions, namely the *restrictor* and the *matrix* (see definition in (40), taken from Krifka et al. (1995), p. 26).

- (39) DEFINITION \mathbf{Gn} is an operator which changes a particular predicate to a characterizing one. Whenever $\mathbf{Gn}(\alpha)(\beta)$ holds, there are several times t and realizations \mathbf{y} of β , $\mathbf{R}(\mathbf{y},\beta)$, such that $\alpha(\mathbf{y})$ holds at t . (with α being a verbal predicate and β a term denoting an *individual* (i.e. a kind or object); realizations may include *stages* of individuals, i.e. temporal slices of an individual, individuals-at-a-certain-time-interval; *John smokes*, for instance, is represented as $\mathbf{Gn}(\text{smoke})(\text{John})$)
- (40) DEFINITION $\mathbf{Q}[\mathbf{x}_1, \dots, \mathbf{x}_i; \mathbf{y}_1, \dots, \mathbf{y}_j](\mathbf{Restrictor}[\mathbf{x}_1, \dots, \mathbf{x}_i]; \mathbf{Matrix}[\{\mathbf{x}_1\}, \dots, \{\mathbf{x}_i\}, \mathbf{y}_1, \dots, \mathbf{y}_j])$, \mathbf{Q} is a dyadic adverbial quantifier; $\mathbf{x}_1, \dots, \mathbf{x}_i$ are the variables to be bound by \mathbf{Q} , and $\mathbf{y}_1, \dots, \mathbf{y}_j$ are the variables to be bound existentially with scope just in the matrix.

Depending on the analysis one chooses, expressions may (but do not have to) have scope over other expressions. Thus, the choice on the specificity of one expression has an influence on the interpretation options of following expressions.

Regarding Example (41a), one option is to interpret both expressions *companies* as intensional. This interpretation would read as ‘It is a typical trait of objects of the type *company* that they buy (or are bought by) other objects of the same type, and this will carry on in the future’. Another option is to interpret both as specific: there will be some companies which buy other companies. A third option is to interpret the first as intensional and the second as specific. In those cases where the first expression is interpreted generically, the second expression is under the scope of this expression. It is not possible to interpret the first expression as specific and the second as generic.

Example (41b) can be paraphrased as ‘the total number of genes identified is 2’. As to truth conditions, it is not necessary that both of these genes have been identified by one and the same scientist (or group of scientists).

- (41) a. Companies₁ are still going to buy companies₂ around the world.
 b. To date, scientists have fingered two of these cancer-suppressors.

		NP ₁ (antecedent candidate)	
		specific	generic
NP ₂	specific	NP ₂ def: coref	¬coref
		NP ₂ indef: ¬coref	
	generic	¬coref	coref*

Table 2.2: Specificity, Genericity and Coreference (*under the assumption that generic NPs refer to a property or to all individuals of the respective kind)

One acknowledged difficulty with generic uses of expressions is that they are sometimes hard to distinguish from specific reference (consider the examples in (42)): Carlson (1989) notices a lack of formal features, Herbelot’s (2008) annotation scheme for genericity consists of a 14-step decision tree. What is more, a text can implicitly restrict the class of objects of interest. For instance, consider the text in (33): the expression *older bridges* refers to older bridges in general, but only those located in the U.S., not in Canada etc.

- (42) a. Mobil Corp. is preparing to slash the size of its work force in the U.S., possibly as soon as next month. [...] Employees haven’t yet been notified.
b. [...] developers may have to put in a lot of money and time.
c. [...] political dissidents were being certified as insane
d. Successful American business owners do the same thing.

Distinguishing specifically referring expressions, however, is crucial for coreference resolution because of the asymmetry in Table 2.2²⁷: whereas coreference is *not* possible between NPs of distinctive types (one specific, one generic), and possible between generics (see Example (43)²⁸) and between definites, it is not possible between specific indefinites.

- (43) Sheep₁ (*Ovis aries*) are quadrupedal, ruminant mammals typically kept as livestock. Like all ruminants, sheep₁ are members of the order Artiodactyla, the even-toed ungulates. Although the name “sheep” applies to many species in the genus *Ovis*, in everyday usage it almost always refers to *Ovis aries*. Numbering a little over one billion, domestic sheep are also the most numerous species of sheep₁. Sheep₁ are most likely descended from the wild mouflon of Europe and Asia. One of the earliest animals to be domesticated for agricultural purposes, sheep₁ are raised for fleece, meat (lamb, hogget or mutton) and milk. A sheep₁’s wool is the most widely used animal fiber, and is usually harvested by shearing. Ovine meat is called lamb when from younger animals and mutton when from older ones. Sheep₁ continue to be important for wool and meat today, and are also occasionally raised for pelts, as dairy animals, or as model organisms for science.

²⁷Abbreviations: coref - coreference is possible (¬coref - coreference is not possible, def - syntactically definite, indef - syntactically indefinite.

²⁸Example taken from <http://en.wikipedia.org/wiki/Sheep>, last access January 29th, 2011.

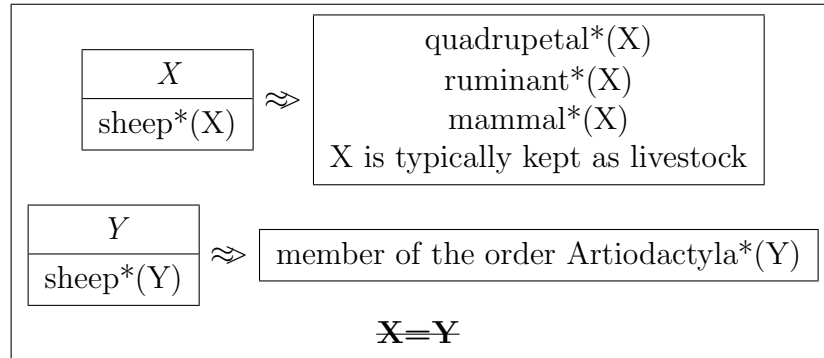


Figure 2.10: DRS of ‘Sheep are quadrupedal, ruminant mammals typically kept as livestock. Sheep are members of the order Artiodactyla.’

Coreference between generic expressions is not provided for in DRT: identity of the variables X and Y ($X = Y$) cannot be postulated because the variables X and Y only have scope inside the respective DRSs, see Figure 2.10. This is the case even if the generic operator (\approx) is substituted by a universal quantifier (\Rightarrow), which is a plausible analysis in the example case. In the DRT analysis, the variables are identical only in extension.

2.2.6 Vagueness and Ambiguity

Unique identifiability, i.e. the clear delimitation of an expression’s referent, is difficult if the text is not explicit as to

- (i) the identity, or the exact extension of the referent or
- (ii) whether it refers to the intension or extension of the expression.

If the speaker does not specify the identity or exact extension of the referent, identity (here: sameness) or overlap with referents mentioned earlier or later in the text is possible, but not specified.

Expressions that do not signal unique identifiability of the referent will be termed *vague* here. In particular, this is the case with interrogatives and expressions denoting uncountable items or abstract concepts rather than concrete objects. Expressions with several possible referents are called *ambiguous*.

Regarding interrogative constituents, i.e. constituents containing interrogative determiners, pro-forms or pro-adverbs may be used to introduce or narrow down a discourse referent, presupposing the existence of a specific referent, as in Examples (44a) and (44b)). Their purpose, however, is to explicitly ask for or leave open the extension of a certain discourse referent (cf. Krifka (2008)), thus the referent is not uniquely identified. It is debatable whether they refer at all.

- (44) a. What must your management team do to establish corporate due process?

- b. Dr. Knudson [...] assumed the missing piece contained a gene or genes whose loss had a critical role in setting off the cancer. But he didn't know which gene or genes had disappeared.

The exact extension of the referent is hard to define for uncountables (Examples (45) to (47)). This also holds for immaterial goods (abstract concepts like *peace* in Example (48)²⁹). They are representations existing in the mind rather than in the real world, which leads to a conflict with the strict definition of referentiality.

- (45) The Valujet statement did not mention smoke in the cabin, but other reports said the pilot had spoken to air-traffic controllers of smoke in the cabin.
- (46) Santa Fe International Corp. [...] is stepping up development of a well off Texas' Matagorda Island where it found gas in 1987.
- (47) Businesses "want to verify information and ensure accuracy," says John Hiltunen [...].
- (48) "[...] Blessed indeed are our friends and colleagues who perished on a mission of peace." [...] Galbraith said the most fitting memorial would be to make peace "a reality on the ground" and to offer "a vastly better future to people who in the last five years have suffered so much."

Vagueness is a very common issue: in principle, any indefinite expression is vague considering that names could be used instead (see Examples (49) and (50)³⁰).

- (49) GE Chairman John Welch has been "besieged with phone calls" [...], according to a person close to him.
- (50) Benson said investigators are simulating air loads on the 737's rudder.

Finally, expressions can be *ambiguous* with respect to their type (specific/non-specific/generic), or their referent (multiple possible antecedents).

Karttunen notes that "[i]n general, indefinite noun phrases have both a specific and non-specific interpretation" (Karttunen, 1969a, p. 6). Lyons supposes the "vagueness between specific and non-specific is in principle always present" (Lyons, 1999, p. 173). Examples for ambiguities between specific/generic uses of expressions are given in (42).

For an example with multiple possible antecedents, see Example (51) (taken from Poesio and Artstein (2005), p. 76, linebreaks changed and markup added), where *that* might refer to *a bad wheel* or *the boxcar*.

- (51) it turns out that the boxcar at Elmira₁ has a bad wheel₂ and they're .. gonna start fixing that_{1 or 2} at midnight but it won't be ready until 8

To conclude Sections 2.2.4 to 2.2.6, the idea behind the definition of specificity was to break down a task into smaller subtasks, namely to identify those expressions for which coreference is clearly and efficiently decidable. The corpus data suggests that non-specificity of expressions is a problem indeed, but deciding on specificity is, too.

To sum up the definition of reference (Section 2.2), being able to refer is a property primarily associated with noun phrases (cf. also Webber (1988)); exceptions to this rule

²⁹Examples (45) and (48) are taken from the MUC-7 corpus; (46) and (47) from OntoNotes 1.0.

³⁰This example is taken from the MUC-7 corpus.

have been discussed above. Basically, an expression is referring in the strict sense if it is used to select a concrete object (or set of objects) (van Deemter and Kibble, 2000), see the definition in (14).

2.3 Coreference

Coreference is a relation between expressions where identity holds between the referents of these expressions (Definition (52) taken from van Deemter and Kibble (2000), p. 629).³¹

- (52) DEFINITION
 $C(\alpha_1, \alpha_2)$, i.e. α_1 and α_2 corefer if and only if $\text{Ref}(\alpha_1) = \text{Ref}(\alpha_2)$, where $\text{Ref}(\alpha)$ is short for ‘the entity referred to by α ’

In logics, identity between objects (denoted by the equals sign ‘=’) is defined as given in (53) (Reinhardt and Soeder, 1974, p. 19): with respect to every property P , the objects are equal in truth values.

- (53) DEFINITION
 $x = y : \Leftrightarrow \forall P(P(x) \Leftrightarrow P(y))$

Identity is an equivalence relation (cf. the discussion in Kibble and van Deemter (2000)). An equivalence relation is reflexive, symmetric, and transitive (see Definitions (54) to (56); cf. Meschkowski (1966), Behnke et al. (1958), Hasse (1963)).

- (54) DEFINITION: reflexivity
 $\forall \alpha \in X, C(\alpha, \alpha)$
- (55) DEFINITION: symmetry
 $\forall \alpha_1, \alpha_2 \in X, C(\alpha_1, \alpha_2) \Rightarrow C(\alpha_2, \alpha_1)$
- (56) DEFINITION: transitivity
 $\forall \alpha_1, \alpha_2, \alpha_3 \in X, C(\alpha_1, \alpha_2) \wedge C(\alpha_2, \alpha_3) \Rightarrow C(\alpha_1, \alpha_3)$

Identity is a complex philosophical concept with many controversial issues (see Noonan (2009) for an overview). Those issues linguistically relevant to this work will be discussed in the following, in particular the issues in determining coreference, and the distinction of coreference from other forms of identity.

³¹In their functional definition $\text{Ref}(\alpha)$, the parameter of the expression’s context is concealed (cf. the discussion in 2.2). Taking the context c into account (and the sentence context might not be enough, consider Examples (36d) to (36e) and the second sentence in (59)), the resulting function $\text{Referent}(\alpha, c)$ might contain circular interdependencies of the sort $\text{Ref}(\alpha_1, c_1) \stackrel{?}{=} \text{Ref}(\alpha_2, c_2)$ where $\alpha_1 \in c_2$ and $\alpha_2 \in c_1$. Considering (36e), let α_1 be *Small neighborhood businesses*, and α_2 be *those businesses*. For resolving what α_1 (*Small neighborhood businesses*) refers to (a specific set, or a kind), at least the whole sentence needs to be taken into account; the sentence includes α_2 (*those businesses*). For resolving what α_2 refers to, again, the whole sentence (and probably the previous discourse) needs to be taken into account; this also includes α_1 .

2.3.1 Consequences of Vagueness, Ambiguity and Non-Specificity

As mentioned in the Section 2.2.6, what an expression refers to can

- (i) be left vague by the speaker,
- (ii) have multiple solutions or
- (iii) be subject to a controversial debate.

In the case of vague expressions, the discourse does not provide details on the expression's exact extension. As a consequence, the recipient is not sufficiently informed to determine whether identity between such expressions holds or not (consider multiple mentions of *investigators* in Example (57)³²); coreference should not be assumed. As to the speaker's motivation for not being more specific, the recipient can only speculate; possibly, he judges the exact relation between the two referents (or groups of referents) irrelevant to the discourse.

- (57) [...] Federal investigators₁ continued their₁ search for the cause of the crash Tuesday, the 20th day since Flight 800 exploded in midair off Long Island and plunged into the Atlantic Ocean, killing all 230 people on board. On the seas and on the shore, investigators₂ said they₂ made a modest amount of progress, though they₂ still have not determined if the plane crash was caused by a bomb, a missile attack or a mechanical malfunction. At the former Grumman hangar in Calverton, investigators₃ on Tuesday began piecing together the fractured parts of the airplane. They₃ also pulled about one-third of the cockpit wreckage off the one-ton ball of metal, essentially unwrapping it.

Ambiguity mainly challenges the representation of the coreference annotation. Example (58) (taken from Poesio and Artstein (2005), p. 76, and in parts discussed above) shows an utterance with two ambiguous expressions, *that* and *it*: By *that*, the speaker can refer to *the boxcar at Elmira* or *a bad wheel*. By *it*, he can refer to *the boxcar at Elmira*; if the task is to find the closest antecedent, *that* is also a candidate, but only in its reading as referring to the boxcar. In (58a and b), the analyses are given.³³ (58c) gives an indexing annotation, where each discourse referent *d* points to the discourse referent first mentioning *d*'s referent.³⁴

- (58) it turns out that the boxcar at Elmira₁ has a bad wheel₂ and they're .. gonna start fixing that₃ at midnight but it₄ won't be ready until 8
a. {1,3,4} (the boxcar, that, it)

³²This example is taken from MUC-7.

³³The numbers correspond to the indices used in the example, the sets represent equivalence sets. Analyses b. and c. are 'packed'; the pipe character '|' stands for *or*, i.e. '{1|2}' is short for '1 or 2'. It is important that 'or' is not interpreted as 'and' (set union), otherwise transitivity and symmetry are not preserved.

³⁴Note that if a discourse referent *d* is annotated with the ID of the closest preceding coreferent expression, the antecedent's ambiguities are 'inherited': *it* in Example (58), for instance, would be annotated with '1|3', and the dispreferred analysis {2,3,4} referring to *a bad wheel* would arise.

- b. {1,4} (the boxcar, it) and {2,3} (the wheel, that)
- c. it turns out that the boxcar at Elmira₁ has a bad wheel₂ and they're .. gonna start fixing that_{3;coref(1|2)} at midnight but it_{4;coref(1)} won't be ready until 8

Generic ('kind-referring') utterances make statements about types of objects, events, states or situations (*Parents* in Example (59), for instance, refers to *parents in general*). There is controversy, however, whether they should be interpreted as referring to *all* objects etc. of a certain kind, *most*, or *some* (cf. Section 2.2.5), or to the intension. If they refer to the intension or to *all* objects/states/events of this kind, coreference holds between the mentions of this kind (elements indexed 1, 2 and 3 in Example (59)). If they refer to *most* objects/states/events of this kind, their extension is again underspecified, and it is not clearly resolvable whether coreference holds or not.

- (59) Parents₁ should be involved with their₁ children's education at home, not in school. They₁ should see to it that their₁ kids don't play truant; they₁ should make certain that the children spend enough time doing homework; they₁ should scrutinize the report card. Parents₂ are too likely to blame schools for the educational limitations of their₂ children. If parents₃ are dissatisfied with a school, they₃ should have the option of switching to another.

2.3.2 Coreference vs. Lexical, Intensional or Extensional Identity

Coreference can be delimited against

- (i) lexical identity,
- (ii) identity of intension and
- (iii) identity of extension.

(i) Multiple uses of the same signifiers (or of the corresponding pro-forms) in one discourse do not necessarily have the same referent, (cf. Examples (60a)³⁵ and b³⁶). In Example (60a), the second use of *his or her successor* can only be interpreted as referring to the successor's successor. In Example (60b), the expression *producers* is used two times, generically in the first case (referring to *producers in general*, see Section 2.2.5 for a discussion of genericity) and specifically in the second case (referring to *some producers*). Generic and specific uses ('types and tokens' in Hirschman and Chinchor's terms) are distinct in reference. On a related note, Hirschman and Chinchor also call the attention to metonymic expressions. Using Example (60c), they explain that the first use of *the New York Times* refers to a newspaper copy, the second to a publishing house.³⁷

³⁵Text source: <http://www.oi.edu.pl/old/php/ceoi2004.php?module=show&file=history>, last called August 9th, 2011.

³⁶Examples in this section taken from Hirschman and Chinchor (1997), pp. 12 and 13 unless specified. Markup changed and inserted.

³⁷Although the example is questionable (the second use of *the New York Times* can be interpreted as also referring to the newspaper), the point is made clear.

- (60) a. Suppose somebody sends a message to his or her successor, who in turns transmits the message to his or her successor, etc.
- b. producers don't like to see a hit wine increase in price... Producers have seen this market opening up and they're now creating wines that appeal to these people.
- c. I bought the New York Times this morning. I read that the editor of the New York Times is resigning.

(ii) Pronouns can be used as identity-of-sense anaphors, as Karttunen (1969b) points out (consider his example in (61), markup added).

- (61) The man who gave his paycheck to his wife was wiser than the man who gave it to his mistress.

(iii) Uses of different descriptions that apply to the same physical object do not necessarily result in coreferential use of these references: In a discourse, one can refer to different aspects or roles of the same object in the real world, see Examples (62) (Example 62a made-up, (62b) from OntoNotes 1.0). Recasens et al. (2010) call relations between objects and their aspects near-identity (see their Examples (62c) and (62d), p. 151 and 149, respectively, (Recasens et al., 2010)). Objects may take on roles, see Example (62e) from (Naumann, 2006, p. 13; original annotation of discourse referents). The characteristics of personal identity may be referred to as persisting even after death (i.e. after an objects existence in the world has ended), cf. Example (62f).

Whether an individual is identical to stages of this individual (cf. definition of stages in the context of generalizations (39)) is an open question.

- (62) a. He wrote a book on the morning star₁, and another one on the evening star₂. (That was before people knew that the morning star and the evening star are the same object.)
- b. [...] A few months later, Mr. Bush₁ became Ronald Reagan's running mate_{att:1}. Suddenly, George Bush the pro-choice advocate₁ became George Bush the anti-abortionist_{1?}. [...] In addition to supporting the landmark Roe vs. Wade Supreme Court decision legalizing abortion, Mr. Bush₁ said he₁ opposed the constitutional ban on abortion that Mr. Reagan was promising to promote. As Mr. Reagan's running mate, though, Mr. Bush_{1?} plunged headlong into the anti-abortion position [...].
- c. "Your father₁ was the greatest" commented an anonymous old lady while she was shaking Alessandro's hand – Gassman₁'s best known son. "I will miss the actor_{2?}, but I will be lacking my father_{1?} especially," he said.
- d. On homecoming night Postville₁ feels like Hometown, USA, but a look around this town of 2,000 shows it₁'s become a miniature Ellis Island ... For those who prefer the old Postville_{1?}, Mayor John Hyman has a simple answer.
- e. John Travolta₁ as a lawyer from Boston sues two companies which he₁ considers responsible for the death of eight children as a result of leukaemia. In the beginning, the calculating high flying advocate₁ only scents high compensation sums, but slowly the case becomes a selfdestroying obsession. Court drama, environmental thriller and great actor's cinema, where Travolta₁ and his₁ antagonist Robert Duvall achieve top form.

- f. During the Korean War, Gen. Douglas MacArthur₁ demanded and got, in addition to his U.N. command in Korea, his own naval command in Japan, NavforJapan. Those obsolete operations cost less than \$ 2 billion a year, and keep Mac's₁ ghost quiet.

Expressions can be used intensionally (attributively/generically) or extensionally. Whether an expression is used in its intensional or extensional sense is usually not explicitly marked. See also the discussion in Section 2.3.4.

2.3.3 Identity of Binding Index

Heim and Kratzer (1998) define *binding* as follows, distinguishing *syntactic* and *semantic binding*, see (64) (on the basis of (63), both definitions are denoted as ‘standard definitions’) and (65), respectively:

- (63) DEFINITION *C-command*
 “A node α c-commands a node β iff
 (i) neither node dominates the other, and
 (ii) the first branching node dominating α dominates β .” (Heim and Kratzer, 1998, p. 261)
- (64) DEFINITION *Syntactic binding*
 “A node α syntactically binds a node β iff
 (i) α and β are co-indexed,
 (ii) α c-commands β ,
 (iii) α is in an A-position, and
 (iv) α does not c-command any other node which also is co-indexed with β , c-commands β , and is in an A-position.
 ‘A-positions’ are the positions of subjects and objects: ‘non-A (A-bar) positions’ are adjoined and complementizer positions.” (Heim and Kratzer, 1998, p. 261)
- (65) DEFINITION *Semantic binding*
 “A DP α *semantically binds* a DP β (in the derivative sense) iff β and the trace of α are (semantically) bound by the same variable binder. [Above this definition, Heim and Kratzer note:] (On the literal notion, only variable binders in the semantic sense can bind anything.)” (Heim and Kratzer, 1998, p. 263)

Similarly, Büring (2007) observes that the term *binding* has been used for the following three phenomena:

“First, for the relation between quantified expressions and pronouns that referentially depend on them [see his example in (66)]. Second for coreference, the relation between two referring expression with the same referent, [see his examples in (67)] including hypothesized empty pronouns, [see his example in (67c)]. Third, in theories that assume transformations, for the relation between a dislocated phrase and its trace, [see his example in (68)].”³⁸

³⁸No page numbering available.

- (66) *Every cat* chased *its* tail.
- (67) a. *Sue* hopes that *she* won.
 b. *Edgar* spoke for *himself*.
 c. *Wesley* called PRO to apologize.
- (68) a. *Which book* did Kim read *t*?
 b. *Antonia* was promoted *t*.

Büiring elaborates that “[s]emantically, only [(66)] and [(68)] are clear instances of binding”³⁹. Throughout this work, the term *binding* will be used in this stricter sense, excluding coreference.

In contrast to the coreference relation, the binding relation involves a *binder* (the antecedent) and a *bindee* (the bound pronoun or noun phrase) (Büiring (2005), cf. also Heim and Kratzer (1998)). Thus, it is not a symmetric relation (see (55) for a definition of symmetry, cf. the discussion in Kibble and van Deemter (2000)). If it was to be included in coreference, transitivity could no longer be guaranteed as a consequence.

Anaphors bound by quantifiers (Examples (66) and (69a)) or in generic utterances (Examples (69b and c)) are different from specific anaphors as in (69d), where the referents of *this robin* and *it* are identical. In (69a, b and c), the whole utterance makes a proposition about a type, i.e. a set of objects (in this case, *robins*), whereas the embedded proposition (*when [...] hungry*) is made for each of the objects in the set, but not for the set as a whole. Thus, ‘identity of referent’ does not seem to describe the relation between these expressions precisely.⁴⁰

- (69) a. Every robin sings when it is hungry.
 b. The robin sings when it is hungry.
 c. Robins sing when they are hungry.
 d. This robin sings when it is hungry.

Carlson (1977b) showed that “anaphoric bindings are possible across kind[-]referring and apparently object-referring uses” (Krifka, 2004, p. 113), more specifically bindings in either direction (consider Carlson’s examples in (70)), cf. Rooth (1995).⁴¹

- (70) a. Even though my brother hates *snakes_{kind}*, I’ve had *them_{specific}* for pets my whole life.
 b. Bob the hunter killed *buffalo_{specific}* until *they_{kind}* were extinct.

³⁹Example numbers have been changed in this citation to match numbering in this work.

⁴⁰Kamp and Reyle (1993) distinguish what they call collective and distributive readings. Their Example (1) can be read as (a) ‘There is a secretary that the group of lawyers liked, and they hired her’ (collective) or (b) ‘Each of the lawyers hired a secretary he liked’ (distributive). Strictly speaking, ‘*they*’ is coreferent to ‘*The lawyers*’ only the collective reading, ‘*they*’ in the distributive reading stands in an anaphoric relation to ‘*The lawyers*’.

(1) The lawyers hired a secretary they liked.

⁴¹For details on the distinction of reference to kinds vs. reference to objects (also called type/token distinction), see Section 2.3.2.

Even non-referring phrases can be related back to anaphorically. The following Examples (71)⁴² and (72) are taken from TüBa-D/Z. Note the differences between glosses and translations. The expressions *Krieg führen* and *Staub aufwirbeln* cannot be translated compositionally (word by word), they are collocational or idiomatic, respectively. This special case of anaphoric reference may be categorised as a play with words. It is probably found more frequently in commentaries than in news reports. Nevertheless it is relevant to the distinction of coreference from anaphoric relations.

- (71) *Wirklich kalt ist nur der, der den Krieg ebenso führen wie auch auf ihn verzichten kann.*
Really cold is only that who the war just as lead just as from it
refrain can.

The really cold party is the one which can wage war just as it can refrain from it.

- (72) *Und die vor zwei Jahren vom Senat beauftragte Unternehmensberatung McKinsey begutachtete die Bremer Kulturszene und wirbelte damit viel Staub auf. Der hat sich zwar wieder gelegt.*
And the ago two years by the senate hired business consultancy McKinsey surveyed the Bremen cultural scene and swirled with this
much dust up. He has itself indeed again settled.

Management consultants McKinsey did a survey of the Bremen cultural scene, commissioned by the senate two years ago, which caused a great stir. It has calmed down since. [...]

Syntactic rules account for a substantial proportion of relations like pronoun binding (e.g., of reflexive, reciprocal, possessive and relative pronouns) and argument control (identity of arguments represented by traces, see Figure 2.11) (Büring, 2005). These rules are, however, not fully deterministic (cf. Büring (2005)):⁴³ Example (73), taken from Hemforth et al. (2000), p. 266, is ambiguous as to the attachment point of the relative clause *who came from Germany*; both *the teacher* and *the daughter of the teacher* are possible points of attachment. Example (74) (taken from Sag and Pollard (1991), p. 85) is ambiguous with respect to the subject of the embedded clause *to be allowed to attend the reception*: syntactically, both *Jim* and *Mary* are plausible arguments.

- (73) [The daughter of [the teacher]₁]₂ who_{1/2} came from Germany met John.⁴⁴

- (74) Jim_j promised Mary_m [PRO_{j/m} to be allowed to attend the reception].

The need for manual disambiguation of such relations as binding and control (represented by links between nodes and traces) might be a motivation to treat them like coreference;

⁴²The immediate context of Example (71) is as follows: *Die Linke, die heute den Krieg führt, tut dies nur halbherzig. Sie liebäugelt mit der Kälte, aber sie rechtfertigt sich mit der Wärme.* ‘The left wing, which is conducting the war [in Kosovo, NB], is doing it half-heartedly. It is flirting with coldness, but justifying itself with warmth.’

⁴³Markup/traces were added to the examples.

⁴⁴In Hemforth et al. (2000)’s study, this and other examples were presented to participants (in German), who should disambiguate them. In German, all relative clauses are comma separated (in English, only attributive relative clauses are); a distinction between the two categories is not always trivial.

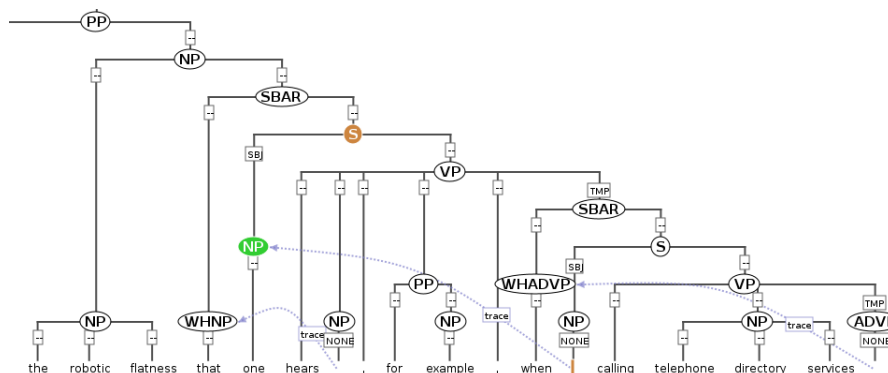


Figure 2.11: Syntactic Tree Containing Traces (OntoNotes 1.0 corpus, visualized in AN-NIS2)

not all binding relations, however, are coreference relations. Theory distinguishes attributive and restrictive relative clauses (cf. Fabb (1990)): In Example (75a), the relative clause gives additional information on a referent mentioned by the name *Robert Wussler*; the relative pronoun is (in the strict sense) coreferent with the named entity expression. In (75b), the relative clause restricts the set of *persons* to the subset of *persons with a defective or early-damaged copy of a suppressor gene*; strictly speaking, the relative pronoun is bound.⁴⁵

- (75) a. Comsat Video is headed by Robert Wussler, *who* resigned his No. 2 executive post with Turner Broadcasting System Inc. just two weeks ago to take the Comsat position.
- b. A person *who* is born with one defective copy of a suppressor gene, or in whom one copy is damaged early in life, is especially prone to cancer because he need only lose the other copy for a cancer to develop.

As discussed in Section 2.2.1, categorizing relative pronouns as mentions of referents is not desirable. This also applies to reflexive pronouns that the verb requires. Reflexive verbs include e.g. *to enjoy oneself*, *to perjure oneself* in English; *sich konzentrieren* - ‘to concentrate’, *sich verlieben* - ‘to fall in love’, *sich erinnern* - ‘to remember’, *sich freuen* - ‘to be delighted’ etc. in German.

2.3.4 Asserted Identity

Identity can be postulated explicitly. Kamp and Reyle (1993), rather in passing, distinguish stipulated from asserted coreference (Examples from Kamp and Reyle (1993), pp. 37 and 257, indices added; cf. also Büring (2005) ‘*identity statements*’ (p. 155)): in Example (76a), the recipient infers the identity between *Jones* and *him*, and between

⁴⁵Kamp and Reyle (1993) suggest that nouns with their respective restrictive relative clauses should be “treated as ‘complex nouns’” (p. 46), using only one variable for all restriction predicates. In that respect, all relative pronouns, whether bound or coreferent are treated equally.

Ulysses and *it* (represented by co-indexing in the examples), respectively. This involves a process of “interpretative *stipulation*, to the effect that a certain anaphoric NP is being used as a means for picking up a certain element introduced into the discourse by independent means” (p. 260). In Example (76b), in contrast, the identity between *Fred* and *the manager of Silver Griffin* is asserted, i.e. stated explicitly.

Kamp and Reyle interpret the constituent *the manager of Silver Griffin* as referring. They suggest representing asserted identity as ‘*x is y*’ (as opposed to ‘*x = y*’ for stipulated identity), though with the same truth conditions, arguing that the assertion’s “particular *contribution* [...] to the interpreter’s information” (Kamp and Reyle, 1993, p. 259) should be made explicit.⁴⁶ Thus, *x is y* and *x = y* are semantically equivalent, but pragmatically different.

- (76) a. Jones₁ owns Ulysses₂. It₂ fascinates him₁.
 b. Fred₁ is the manager of Silver Griffin_{=1|att:1}.

However, Kamp and Reyle also note that “[n]ot all uses of the verb **be** express identity” (p. 260). They call this other use, which “attribut[es] properties to individuals[,] [...] *predicational* use” (Kamp and Reyle, 1993, p. 260), see their example in (77), with the predicate printed in italics.

- (77) John is *happy*.

The construction in Example (76b) is also possible without the definite determiner in the predicate, see Example (78a)⁴⁷ (and Example (78b) for additional evidence from the OntoNotes corpus).

- (78) a. Fred is *manager of Silver Griffin*.
 b. Mr. Tucker, 44 years old, is *president of Trivest Securities Corp.*

While the interpretation as a referring expression seems plausible for a definite description like *the manager of Silver Griffin*, such an interpretation seems less natural if the determiner is omitted: the corresponding indefinite singular is not commonly used referringly (see Examples (79a) and (79b)). This holds for English and German.

- (79) a. #*Manager of Silver Griffin* announced a new strategy yesterday.
 b. #Fred met *manager of Silver Griffin/president of Trivest Securities Corp.*

⁴⁶Cf. Frege’s (1892) argumentation that sentences of the form *a=b* often ‘valuably increase our insight’ (as compared to sentences of the form *a=a*, i.e. tautologies). Considering Washburn’s (1898) observations below, it seems that such sentences have definitional character, and are most adequate if at least one of the objects or concepts involved are already part of the hearer’s knowledge, providing him with an anchor point for storing further information.

“What mental process is associated with the word ‘is’ or ‘are’? What is the state of consciousness when we declare that A *is* B? [She later gives the example ‘Horses are quadrupeds’, NB.] The fundamental process of mind involved in judgment would seem to be the process by which in a complex conscious state a certain element is fixed upon, analysed out, by the attention, and thus given a greater clearness in consciousness than it had before.” (Washburn, 1898, p. 526).

⁴⁷In contrast to Example (76b), Example (78a) lacks the connotation that Fred is the *only* manager of Silver Griffin.

I interpret this as an argument for an analysis of identity assertions as predicational rather than referential. A referential analysis of the predicates is arguable for purposes of information extraction, where *all* the information available in the text needs to be assigned to the referent it holds for. As noted above, there is no difference between a predicational and a referential analysis from the point of truth conditions; the difference is only on the textual and pragmatic levels. An annotation of referentiality could be applied for excluding non-referring expressions as possible antecedents in systems for anaphora and coreference resolution.

A distinction of predicational and referential uses is not always obvious even for definite descriptions, consider the examples in (80).⁴⁸

- (80) a. The winner is Max.
 b. Max is the winner.

Example (80a) can be argued as presupposing two discourse referents, someone who won, and a person named *Max* (who happen to be identical). It cannot be understood as a predication. Example (80b), however, which is equivalent to (80a) in terms of truth conditions, can be used both as a statement asserting identity between two referents (the person who won, and the person named Max) and as a predication (ascribing the predicate *winner* to the person *Max*, while additionally implying there is only one such *winner*). This is relevant to the incremental processing of discourse: Kamp and Reyle (1993) note that in cases like (76b) and (80b), the first processing step would introduce two distinct variables, e.g. *x* and *y*. In the second step, on processing '*x is y*', *y* could be replaced by *x*, and could be dropped from the universe. The difference between the example is in the topic-comment structure (see Section 1.3).

Predications involving generics describe an asymmetric (directed) relation rather than a symmetric one, cf. Examples (81). Their logic representation is of the form $\forall x(A(x) \rightarrow B(x))$, i.e. B follows from A; the predicates A and B are not commutable.⁴⁹

- (81) a. Whales are mammals.
 b. #Mammals are whales.
 c. The raven is a bird.
 d. #The bird is a raven.

Appositions, which can be interpreted as abbreviated relative clauses '*x*, (who/which is) *Y*, ...', can also be grouped in the category of asserted identity (see Example (82)).

- (82) Earlier this year, Cordis, a maker of medical devices, agreed to plead guilty to felony and misdemeanor charges [...].

Identity assertions can also be made using copula like *seem* and *become* (see Example (83), repeated from (62b) above).

⁴⁸The minimal pair (a, b) is a made up example.

⁴⁹In German, the predicative noun can be fronted, e.g. as a reply to a clarification question like *What are whales?:* SAUgetiere sind sie. (engl. MAMmals are they, i.e. 'Mammals is what they are.'). Repeating the word 'whales' in the answer, however, makes the sentence sound unnatural. Capital letters in the examples represent stressed syllables.

- (83) [...] A few months later, Mr. Bush₁ became Ronald Reagan's running mate_{att:1}. Suddenly, George Bush the pro-choice advocate₁ became George Bush the anti-abortionist_{1?}. [...] In addition to supporting the landmark Roe vs. Wade Supreme Court decision legalizing abortion, Mr. Bush₁ said he₁ opposed the constitutional ban on abortion that Mr. Reagan was promising to promote. As Mr. Reagan's running mate, though, Mr. Bush_{1?} plunged headlong into the anti-abortion position [...].

It is not clear whether constructions with 'as' (*As Mr. Reagan's running mate*) should be treated as identity assertions. They open a distinction of aspects inherent to an entity that has previously been treated as one (and might be continuously treated so).

2.3.5 Identity and Time Dependence

Just as it is valid to use different descriptors for the same referent, the validity or appropriateness of a descriptor may change over time. Reconsider Examples (62) (Section 2.3.2). Kibble and van Deemter (2000) remark that disregarding time dependence may result in wrong conclusions on identity, which challenges the transitivity property of coreference. They argue that in Example (84a)⁵⁰, from identity of referent between *Henry Higgins* and *sales director of Sudsy Soaps*, and between *Henry Higgins* and *president of Dreamy Detergents*, it would follow that identity of referent holds between *sales director of Sudsy Soaps* and *president of Dreamy Detergents* (which is not the case at any point in time). Similarly, the function value of *the stock price* changes over time, thus, this expression has different 'referents' (values) over time (again, these are not identical at any point in time).

- (84) a. Henry Higgins₁, who was formerly sales director of Sudsy Soaps₁, became president of Dreamy Detergents₁.
 b. The stock price₁ fell from \$4.02₁ to \$3.85_{1?}. Later that day, it fell to an even lower value, at \$3.82_{1?}.

It should be noted, however, that from identity of reference, identity of intension does not follow.

2.3.6 Accommodation of Referents

Anaphoric expressions can refer back to sets of referents, to events, states or propositions. These may have been mentioned in other forms than *one* discourse referent, a nominal constituent, see the examples in (85). The hearer has to reorganize the representation of the information already given to get ready for 'docking' the new information.

As to a grouping of previously mentioned referents, Kamp and Reyle (1993, p. 307ff.) propose an operation called *summation*: a set variable is introduced, which groups together the variables it contains, e.g. $Z = u \oplus v$ for Example (85a), where *u* and *v* are the variables *John* and *Mary* are mapped to (and analogously for Example (85b)). This set variable can later be set equal with the variable *they* is mapped to (see DRS in Figure 2.12).

⁵⁰Both examples in (84) taken from Kibble and van Deemter (2000), p. 632f. Markup added.

In the case of quantifiers involved (Example (85c), they suggest an *abstraction* operation: a set variable is introduced that contains all the elements fulfilling the constraints specified in the first sentence. Again, this set variable can later be set equal with the variable *they* is mapped to, see Figure 2.2.

- (85) a. John₁ took Mary₂ to Acapulco. They_{1⊕2} had a lousy time.
 b. Last month John₁ took Mary₂ to Acapulco. Fred₃ and Suzie₄ were already there. The next morning they_{1⊕2⊕3⊕4} set off on their sailing trip.
 c. Susan has found {every book/most books} which Bill needs₁. They₁ are on his desk.

x	y	z	X	Y
John(x)	Mary(y)	Acapulco(z)	x took y to z	X=x⊕y
				Y=X
				Y had a lousy time

Figure 2.12: DRS for ‘John took Mary to Acapulco. They had a lousy time.’

As to other complex antecedents like events and propositions (see Examples (86)) and their representation in DRT, see Asher’s example in Figure 2.13 (Asher, 1993, p. 91).

- (86) a. Weatherford International Inc. said it canceled plans for a preferred-stock swap [...]. Weatherford said market conditions led to the cancellation of the planned exchange.
 b. Reducing those rates moderately [...] would still provide substantial assistance to borrowers. But it would also encourage lenders to choose more creditworthy customers [...].
 c. I saw what he did to them firsthand. It made my shoelaces dance with terror.
 d. In the neighbourhoods with the highest crime rates, small business generally relies on the public police force for protection.
This creates several problems.

2.3.7 Bridging

Besides identity, a referent can be related to the context in many different ways. This causes referents to be implicitly given (from the speaker’s perspective) or accomodated (from the hearer’s perspective) (Clark (1975), also see Poesio and Vieira (1998) for an overview).⁵¹ Such relations include set relations (subset, see Example (87), superset,

⁵¹All examples in this subsection are taken from Clark (1975) unless specified otherwise.

$Y e Z z e'$
students(Y)
e-go-camping(Y)
e'-enjoy(Z,z)
Z=Y
z=e

Figure 2.13: DRS for ‘The students went camping. They enjoyed it.’

member of the same set), the part-whole relation (88), and the entity-attribute relation (89), as well as typical role fillers (“Often the [g]iven information characterizes a role that something implicitly plays in an event or circumstance mentioned before” (Clark, 1975, p. 171), see Example 90). Clark not only extends his definition to *necessary* parts, roles, etc. (Examples (88a), (89a), (90a)), but also to *probable* parts, roles etc. (Examples (88b), (89b), (90b)).

- (87) I met two people₁ yesterday. The woman_{2; 2_⊆1} told me a story.
- (88) a. I walked into the room₁. The ceiling_{2; part-of 1} was very high.
b. I walked into the room₁. The windows_{2; possible-part-of 1} looked out to the bay.
- (89) a. I ran two miles the other day₁. It₁ did me good.
b. I ran two miles the other day₁. The whole stupid business_{2 att:1} bored me.
- (90) a. John was murdered yesterday₁. The murderer_{2; 2 role-in 1} got away.
b. John died yesterday₁. The murderer_{2; 2 possible role-in 1} got away.

In summary, expressions α_1 and α_2 are considered coreferent in terms of DRT if and only if α_1 has introduced a variable into the main DRS which the variable introduced by α_2 can be linked to via the identity relation (‘=’, see definition (52)). Inferring whether or not identity between referring expressions holds is easier the more specific these expressions are. A general issue of coreference is that it is defined as a relation between *variables at the discourse representation level* but is annotated as a relation between *expressions at the linguistic level*. As discussed in Section 2.2.5, expressions on the linguistic level represent abstractions from the world. The more abstract the expression, the more difficult is the decision on the identity of referents.

DRT is a means of representing coreference, but does not provide resolution strategies. I suggest interpreting the conditions specified in DRSs (initially without the coreference conditions) as a filter. Reconsider Example (91) (repeated from (4), but without prosodic information).

- (91) John₁ has an old cottage₂. Last year, he₁ reconstructed the shed_{2?}.

Suppose we have resolved the pronoun *he* as referring to the same entity as the male name *John* on the basis that both expressions are masculine singular.⁵² Next, we have the predicates *old cottage* and *shed*, and it is up to the reader to decide whether they are used to describe the same object in the world. One factor that could give clues on coreference is the role the objects play in relations: the discourse contains a possession relation between *John* and *the old cottage*, and a reconstruction relation between *John* and *the shed*. In this case, it is just as likely that the object reconstructed by its owner is the object of possession, than that it is a part of that object (namely, *the shed*, as opposed to the main building). Yet, if the reader considers it impossible that $\llbracket \textit{old cottage} \rrbracket_{w,t}$ and $\llbracket \textit{shed} \rrbracket_{w,t}$ intersect, he will infer two separate referents.⁵³ He might be inclined to infer coreference between these expressions to the extent to which he considers the extensions of the two predicates likely to intersect. Thus, knowledge on the coincidence of properties might be another factor used to resolve coreference.

Referents of deictic and generic expressions can be accommodated at any point in the discourse. Coreferent or not, they can be interpreted independently. Creating relations between the respective entities is not necessary in terms of truth conditions.

2.4 Coreference, Discourse-Givness and Information Status

Riester (2009) observes that “in most [...] [annotation schemes], discourse givenness is understood as being equivalent to coreference” (p. 147). Following Riester (2009), for this work I define an expression α as *discourse-given* in a context c if it has a coreferent antecedent in that part of c to the left of itself. A formalisation is given in (92).

- (92) DEFINITION
 $\exists(\alpha_1) (\text{Ref}(\alpha_1, c) = \text{Ref}(\alpha_2, c) \ \& \ \alpha_1 < \alpha_2)$; $\alpha_1 < \alpha_2$ stands for ‘ α_1 occurs before α_2 ’; optional additional condition $\text{NP}(\alpha_1)$, i.e. α_1 is an NP

Annotating discourse-givness does not involve specifying which expression is the antecedent. Thus, resolving ambiguities of referent such as *that* and *it* in Example (93) (repeated from (58)) are avoided. Discourse-givness, as opposed to coreference, is a relation of an expression to its context, rather than a relation between two expressions.

- (93) it turns out that the boxcar at Elmira has a bad wheel and they’re .. gonna start fixing **that** at midnight but **it** won’t be ready until 8

In the operationalization used in this work, the concept of discourse-givness is dependent on the definition of coreference, which is different across the resources used here.

Categorizations subsuming discourse-givness include Prince’s (1981; 1992) hierarchical scheme of familiarity, Gundel et al.’s (1993) givenness hierarchy, and Calhoun et al.’s (2005) and Götze et al.’s (2007) information status. Besides discourse-givness, these schemes include categories for discourse referents related to the context via bridging (see

⁵²The fact that both expressions occur as subjects and agents in the respective sentences is further evidence for a coreference relation between them.

⁵³See Section 2.1.2 for the definition of the notation.

Section 2.3.7). These categories are termed *inferable*, *mediated* or *accessible*, respectively. They also take into account the extralinguistic context (the specific utterance situation; some of them also consider frames, i.e. prototypical situations).

Figure 2.14 gives an overview of Prince’s categories.

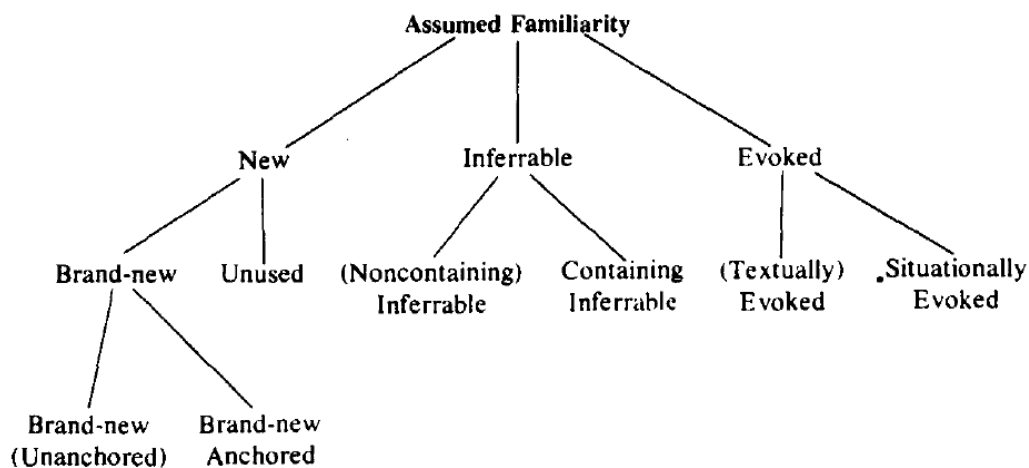


Figure 2.14: Prince’s Hierarchical Scheme of Familiarity (taken from Prince 1981, p. 237).

Prince gives the following examples for her categories (Prince, 1981, p. 233):⁵⁴

- (94) a. I got on a bus*brand-new unanchored* yesterday and the driver*noncontaining inferable* was drunk.
 b. A guy I work with*brand-new anchored* says he*textually evoked* knows your sister.
 c. Noam Chomsky*unused* went to Penn.
 d. Hey, one of these eggs*containing inferable* is broken!
 e. Pardon, would you*situationally evoked* have change of a quarter?

Under the term *cognitive status*, Gundel et al. (1993) distinguish the categories *type identifiable*, *referential*, *uniquely identifiable*, *familiar*, *activated* and *in focus*. This distinction is also called *givenness hierarchy*: the strongest category is *in focus*, an expression *in focus* is also *activated*, *familiar* and so on. They give the following examples (p. 442f.):⁵⁵

- (95) a. I couldn’t sleep last night. {A/This} dog next door*type identifiable* kept me awake.
 b. I couldn’t sleep last night The dog next door*uniquely identifiable* kept me awake.
 c. I couldn’t sleep last night That dog next door*familiar* kept me awake. [“the addressee already knows that the speaker’s neighbor has a dog”, p. 443]
 d. I couldn’t sleep last night. That*activated* kept me awake. [“if a dog has actually been barking during the speech event or if barking had been introduced in the immediate context”, p. 443]

⁵⁴Annotations added as specified in Prince (1981).

⁵⁵Annotations added as specified in Gundel et al. (1993).

- e. My neighbor’s Bull Mastiff bit George. {It’s/That’s}_{in focus} the same dog that bit Mary Ben.

Calhoun et al. (2005) distinguish three categories of information status: discourse referents which are neither previously introduced nor inferable by the hearer from the discourse are labeled *new*. Discourse referents generally known (such as *the moon* etc.), related to the context via possessives, aggregation or bridging (set- or part-whole relation, situationally, or roles in an event), bound pronouns and function values are labeled *mediated*. Generic pronouns, pronouns referring to the dialogue participants, and discourse referents previously mentioned are labeled *old*. The scheme is also hierarchical: each of the categories has subcategories, e.g. *old/ident* for a previously mentioned discourse referent. Götze et al. (2007) suggest a similar distinction: the category *new* is for discourse referents neither mentioned nor inferable. The category *accessible* is for discourse referents that are situationally available (as part of the discourse situation), generally known (as part of the hearer’s world knowledge), aggregated or inferable (via part-whole, entity-attribute, or set relations). The category *given* is for discourse referents that have been explicitly mentioned in the previous discourse. This scheme is also organized hierarchically. See Figure 2.15 (adapted from Götze et al. (2007)) for an overview of the categories and their relation to other concepts.

discourse status	discourse-old	discourse-new	
hearer status	hearer-old		hearer-new
information status	given	accessible/mediated/inferable	new

Figure 2.15: Information Status, Discourse Status, and Hearer Status (taken from Götze et al. 2007, p. 151)

Commonly, the categories of information status are interpreted as a scale, i.e. if an expression is, in a certain context, given and accessible at the same time, the label *discourse-given* ‘overrules’.

With all these categorizations, expressions related to the context by the entity-attribute relation (Example (89b)) are classified as given rather than as accessible/mediated/inferable, which would follow from Clark’s (1975) definition of bridging.⁵⁶ To sum up this section, discourse-giveness and information status are a categories which qualify the relation between a referring expression and its context: this expression is given if the referent of this expression has been mentioned in the context (if it is coreferent to any expression in the context). It is accessible (other terms for this category include mediated or inferable) if it is related to the context by set relations etc.; otherwise it is new.

⁵⁶Clark proposes a definition of givenness extending to other units of discourse (rather than just NPs), see his example in (1) (Clark, 1975, p. 173). So does Schwarzschild (1999), see his definition in (2). These definitions, however, are beyond the scope of this work.

- (1) Alex went to a party last night. He’s going to get drunk again tonight.
- (2) “An utterance U is *GIVEN* iff it has a *salient* antecedent A and A entails U, modulo \exists -type shifting.” (Type shifting “raises expressions to type t, by existentially binding unfilled arguments”; both citations originate from Schwarzschild (1999), p. 146).

Concepts and definitions to remember from this section are shown in Table 2.3. They will be reused in the comparison of annotation schemes in Section 3.2.2.

non-referring	expressions that do not contribute to the sentence's meaning in a fully compositional way (expletives, predications, collocational and idiomatic expressions)
antecedent	an expression preceding the current expression, with a coreference or bridging relation holding between these expressions
binding	a relation between expressions where the second expression is dependent on the first expression (the first expression being quantified or representing a generalization)
event anaphor	an expression referring back to an event previously mentioned using a VP (instead of an NP)
proposition anaphor	an expression referring back to a proposition (not an NP)
predication	a construction explicitly attributing a certain property to the subject NP, e.g. <i>Max is an actor: actor(Max)</i>
asserted identity	a construction stating equality between two objects (or groups of objects), e.g. <i>Max is the winner.</i>
function-value	relation between an expression stating an attribute (e.g. <i>price</i>) and an expression stating its value (e.g. <i>\$ 3.99</i>)
apposition	a construction attributing a certain property to a referent, e.g. <i>Max, the famous actor, won the Oscar.</i>
kind-referring	property of an expression which, in its context, relates to a type of objects rather than a concrete set of these objects (e.g. <i>Dogs</i> in <i>Dogs are my favourite pets</i> as opposed to the concrete set of dogs in <i>Dogs devastated my garden last night.</i>)
abstract concept	immaterial goods; complement to concrete (sets of) objects
specificity/definiteness	property of expressions that refer to concrete (sets of) objects; the expression is used to single out this object (or these objects, respectively)
bridging	relation between expressions, where the expressions' referents stand in a set relation (subset, superset, member of, etc.) or part-whole relation to each other

Table 2.3: Concepts and Definitions

Chapter 3

Corpora and Annotation Schemes

In this chapter, corpora annotated with discourse-givenness, coreference or related concepts are introduced. Their respective annotation schemes are compared to each other, drawing on theoretical definitions presented and discussed in the previous chapter. These corpora form the basis of classifiers for discourse-givenness, which are presented in Chapters 4 (previous work) and 5 (my advancements and furthering experiments).

3.1 Corpora Annotated with Discourse-Givenness or Coreference

In the literature, diverse data sets have been used for training classifiers for discourse-givenness; they will be introduced in turn, characterized by their

1. availability and provider of annotations (publicly available/available on demand/private),
2. source texts (e.g. newspaper texts, dialogues),
3. language and mode (written language, transcribed speech)
4. relevant layers of annotation
5. size (in tokens and/or NPs where available, average discourse length)
6. usage in related work (cf. Chapter 4)

A first overview is given in Table 3.1.

name (year of origin)	language	type of text	docs	NPs	tokens	NPs /doc
MUC-7 (1997)	US English	newswire	53	9,963	30,002	188
Switchboard (2004)	US English	dialogue	147	64,000	(n.a.)	436
OntoNotes 1.0 (2006)	US English	newswire	597	129,781	370,789	217
ARRAU (2008)	US English	mixed	256	62,209	217,485	243
PCC (2004)	German	commentaries	220	4,979	43,652	23
TüBa-D/Z 6.0 (2010)	German	newswire	2,777	373,763	975,836	135
DIRNDL (2012)	German	radio news	55	13,489*	50,000	245

Table 3.1: Overview of Corpora Annotated with Coreference/Discourse-Givenness

Abbreviations: doc - document (news article or dialogue), US - as spoken in the United States

* Of these, approximately 10,000 are referring and thus labeled for information status. The corpus is under construction; the numbers are reported as of February 2013 (personal communication).

3.1.1 MUC-7

The MUC-7 corpus originates from the Message Understanding Conferences (MUC), a series of conferences in the late 1980s and 1990s which issued shared tasks related to information extraction. At MUC-6 and 7 (held in 1995 and 1997, respectively), the shared tasks included coreference resolution. The corpora used in this shared task are available from the Linguistic Data Consortium (LDC)¹. I used the more recent MUC-7 corpus. MUC-7’s primary data consists of New York Times newswire and covers selected topics in aviation and space missions. The language is American English, the mode written language. Distributed as unparsed text, it is manually annotated for coreference according to Hirschman and Chinchor (1997) and named entity types according to Chinchor (1998). The corpus contains appr. 30,000 tokens in total. As shared task data, it is divided into the sets Training, Dryrun and Formaleval for training, dryrun and evaluation, comprising 3,058, 16,018, and 10,926 tokens, respectively. This corresponds to 998, 5,388 and 3,577 NPs after parsing the data with Charniak’s self-trained parser (McClosky et al., 2006).² This corpus has been used for the classification of discourse-givenness by Ng and Cardie (2002) and Uryupina (2003; 2009), and for coreference resolution, for instance, by Soon et al. (2001). MUC-7’s successor in coreference resolution for information extraction applications is ACE (Automatic Context Extraction; Walker et al. (2006)).³ The anaphora resolution system BART (Beautiful Anaphora Resolution Toolkit, Versley et al. (2008b)) has been trained on MUC-6 data.

¹<http://www ldc upenn edu/>, last access 27.03.2013.

²Numbers of tokens and NPs vary across different studies due to tokenization, parsing, and inclusion/exclusion of meta information in article headers (author, news agency, title, short title etc.).

³ACE is geared towards Information Extraction tasks, i.e. coreference is used as a means of collecting all the information on a certain entity available in a text. Its coreference definition is thus focussed on certain entity types, such as person, organization etc. (cf. Uryupina and Poesio (2012)), and extends to predication (this issue is addressed in Section 2.3.4).

3.1.2 Nissim’s Annotation of the Switchboard Corpus

The Switchboard corpus (Godfrey et al., 1992) is a corpus of spontaneous two-party telephone conversations. Nissim et al. (2004; 2006) describe the annotation of information status to this corpus. Whereas the original Switchboard corpus is available via LDC, Nissim’s annotation is not publicly available to my knowledge. The language of the source data is US English, the mode is transcribed speech. The original annotation of Switchboard comprises phonetic and orthographic transcriptions, parts of speech, and speech acts. According to Nissim et al. (2004), a third of the Switchboard corpus is part of the Penn Treebank (Marcus et al., 1993) and thus syntactically annotated according to Penn Treebank conventions. Nissim’s annotations extend over 147 dialogues (approx. 64,000 NPs). This corpus has been used for the classification of discourse-givenness and information status by Nissim (2006) and Rahman and Ng (2011).

3.1.3 OntoNotes

OntoNotes (Hovy et al., 2006; Weischedel et al., 2007), is a corpus of American English, Chinese and Arabic annotated at multiple syntactic and semantic layers. It is available via LDC. In this work, only the English portion of version 1.0 will be used; as from now, the term *OntoNotes* will be used to refer only to this portion unless specified otherwise. OntoNotes consists of 597 documents (370,789 tokens, 129,781 NPs) of newswire from the nonfinancial news portion of the *Wall Street Journal* (WSJ). The English portion of the most recent version 4.0 contains more newswire (now 600,000 tokens), as well as broadcast news (200,000 tokens), broadcast conversation (200,000 tokens) and web text (300,000 tokens).

OntoNotes builds on the Penn Treebank (Marcus et al., 1993) for syntax (including parts of speech) and Penn PropBank for predicate-argument structure, respectively. It contains annotations of named entity types, and word senses are disambiguated with reference to the WordNet ontology (Miller, 1995). It is annotated for coreference according to the OntoNotes coreference guidelines (2007, authorless). Markert et al. (2012) have added annotation for information status according to Nissim et al. (2004), reusing the existing coreference annotation.

The CoNLL (Conference on Computational Natural Language Learning) shared task 2011 of coreference resolution and anaphoric mention detection makes use of the English portion of OntoNotes, version 4.0, as a task specification.

Markert et al. (2012) have built a classifier for information status based on OntoNotes. They used the *Wall Street Journal* section of OntoNotes with its coreference annotation and added annotation of mediated expressions according to Nissim et al. (2004). Uryupina and Poesio (2012) used OntoNotes, along with other corpora, for coreference resolution.

3.1.4 ARRAU

The Anaphora Resolution and Underspecification corpus (ARRAU, Poesio and Artstein (2008)) builds on several existing resources: Trains-91 (Gross et al., 1993), Trains-93 (Heeman and Allen, 1995), GNOME (Poesio, 2004a; Poesio, 2004c), the English Pear Stories corpus (Chafe, 1980), and Penn Treebank (Marcus et al., 1993). It is available on

demand from the authors (provided that licenses to the source corpora are obtained); a release via LDC is planned.

The genres it includes are dialogues (Trains-91, Trains-93), narrative text (GNOME, Pear Stories), and newswire (*Wall Street Journal* portion of the Penn Treebank). The Trains corpora consist of task-oriented dialogues. GNOME consists of pharmaceutical patient information leaflets, informative texts on museum objects, and tutorial dialogues from the Sherlock corpus (Lesgold et al., 1992).

The language is American English. Parts of ARRAU are transcribed speech (Trains-91, Trains-93, Pear Stories, parts of GNOME), other parts are written language (Penn Treebank, parts of GNOME).

ARRAU contains syntax annotation, re-using annotation where existing, otherwise applying Charniak’s (2000) parser and manually correcting its output. Each of the resulting noun phrases is annotated with a set of features including number, gender, person, grammatical function, a feature combining animacy and a concrete/abstract distinction, referentiality, and, if applicable, a link to the expression (or expressions, respectively) it is related to anaphorically or via bridging.

Unlike MUC and OntoNotes, the focus of this resource is ambiguity with respect to the referents of expressions, as well as reference to abstract entities (actions, events, plans). ARRAU consists of 294 discourses, containing 217,485 tokens (62,383 NPs) in total. The corpus has been used for coreference resolution (Uryupina and Poesio, 2012). I am not aware of any works in discourse-givenness classification based on this corpus.

3.1.5 PCC

The Potsdam Commentary Corpus (PCC, (Stede, 2004)⁴) is a corpus of newspaper commentaries from *Märkische Allgemeine Zeitung*, a German regional daily. The corpus is available on demand from the author. The language of the corpus is German, the mode is written language. Annotation includes the following layers: parts of speech according to the Stuttgart Tübingen Tagset (STTS (Schiller et al., 1999), automatically annotated with Brants’ (2000) Trigrams’n’Tags (TnT) tagger⁵), manually annotated syntactic structures according to the TIGER guidelines (Brants et al., 2002) (partially also with TIGER morphology), rhetorical structure according to Rhetorical Structure Theory (Mann and Thompson, 1988), as well as coreference according to the Potsdam Coreference Scheme (PoCoS, (Krasavina and Chiarcos, 2007))⁶ and in parts information structure (information status, topic, focus according to (Götze et al., 2007)). The corpus consists of 43,652 tokens and 4,979 NPs.

3.1.6 TüBa-D/Z

TüBa-D/Z (short for *Tübinger Baubank des Deutschen/Zeitungskorpus*, ‘Tübingen Treebank of German/Newswire Section’) is a German newswire corpus. License is granted

⁴<http://www.ling.uni-potsdam.de/pcc/pcc.html>, last access 27.03.2013.

⁵<http://www.coli.uni-sb.de/~thorsten/tnt>, last access 27.03.2013.

⁶PoCoS consists of a core scheme (comparable to MUC, among others), and an extended scheme (comparable to GNOME). The main differences are (i) only constituents can receive annotations (not parts of constituents, as in MUC), and (ii) heuristics for resolving ambiguity and vagueness.

by Eberhard Karls Universität Tübingen (free academic license).

In this work, version 6.0 of TüBa-D/Z will be used, which consists of nearly 1 million tokens (373,763 NPs, 2,777 articles) from the newspaper *die tageszeitung (taz)* and has been manually annotated with parts of speech (Stuttgart Tübingen Tagset, STTS, (Schiller et al., 1999)), topological fields, syntactic phrase structure trees, and coreference (for documentation, see Telljohann et al. (2006), Naumann (2006)).

TüBa-D/Z has been used for automatic coreference resolution (Versley, 2006), but, to the best of my knowledge, not to classify discourse-givenness.

3.1.7 DIRNDL

DIRNDL (short for Discourse Information Radio News Database for Linguistic analysis, Eckart et al. (2012)) is a corpus of German broadcast news. It consists of transcribed speech. Annotation is ongoing, and licensing conditions are to be defined. The data is parsed with the XLE parser in order to obtain analyses according to Lexical Functional Grammar (Rohrer and Forst, 2006). A subset of around 10,000 referring expressions is manually annotated with information status according to Riester et al. (2007) (Riester et al. (2010) represents an update of the more detailed earlier German version of the annotation guidelines).

Cahill and Riester (2012) have built a classifier for information status in German based on this corpus.

3.2 Comparison of Annotation Schemes

This comparison puts the different annotation schemes in relation to each other and gives an overview of how close they are to the theoretical definition given in Chapter 2.3.

The annotation schemes differ wrt. their definitions of coreference (including preconditions and formalization), and the documentation of annotation quality. A juxtaposition of annotation schemes (in tabular format) will identify where selective comparisons are possible. Comparability of annotation schemes will be of relevance when evaluating different models trained on different resources.

In this section, the schemes will first be compared (irrespective of the fact that resources annotated according to these schemes are available only in certain languages), see Sections 3.2.1 to 3.2.3. Then, decisions of scheme design and the advantages and limitations resulting from them will be discussed (Section 3.2.4).

The annotation schemes differ slightly in their definitions of terms and use of concepts of (1.) referent, (2.) referentiality, (3.) coreference, and (4.) context, as well as (5.) the formalization of the annotation task and (6.) evaluation of the resulting annotation. In particular, they give distinct answers to the following groups of questions:

1. Which linguistic entity can form a markable (i.e. an entity that could be annotated with a label and/or link)?
2. Are there preconditions for coreference (e.g. specificity)? Are the same preconditions applied to antecedent and to anaphor candidates? Which markables cannot enter coreference relations? Is this distinction made explicit?

3. How is coreference characterized and distinguished from, e.g. binding, bridging, etc.? Does the scheme further differentiate types of coreference? If so, on which basis, and are they explicitly labelled? Does the scheme require a linking of coreferent expressions?
4. What type of context (textual or situative) is considered in the annotation?
5. What are the annotation model's structural properties? Does it include links? Are labels attached to nodes or edges (if existing)? What interpretation does the annotation allow for, and to what consequence?
6. How was each of the schemes evaluated? Are the evaluation results comparable? How useful are the annotated corpora with respect to replicability of the annotator's decision and interpretability of the annotation?

3.2.1 Markables and Preconditions

Markables are those units of text that may receive annotation (consisting of a label and/or link). There are two ways of proceeding with the annotation:

- the annotator marks only those expressions that are relevant (i.e. involved in a link, either as antecedent or as anaphor), or
- the annotator is given a set of expressions and has to go through a decision process for each of these expressions (e.g. a decision tree as in Nissim et al. (2004)).

In the latter case, it is common practice to define that only instances of certain syntactic categories are eligible to form markables. This limits the number of candidate instances. It also ensures consistency, especially if the annotation is accomplished by multiple annotators. The aim is to reduce annotation time and cost, while ensuring a high quality standard.

In either case, preconditions may be defined, which are to be checked at the beginning of the annotation process. Markables not meeting these requirements are considered irrelevant and are discarded from further annotation, either because later decisions are not applicable, or because a full specification would substantially complicate the decision process.

The definition of markables and preconditions will later be of relevance to the automatic preprocessing of the data as input to the classification process.

3.2.1.1 Markables

There is consensus across all schemes that noun phrases form markables. There are divergences, however, in conceptions of what exactly constitutes a noun phrase, and whether other elements (such as adverbs, verbs or clauses) should also be admitted.

Further controversial aspects include

- (i) referentiality (i.e. an expression's ability to introduce or take up again a discourse referent). In particular, this is the case with possessive, relative, reflexive/reciprocal pronouns, traces/zero forms, interrogative constituents, temporal, local and numeric expressions, mentions embedded in compounds or complex names,

present participles vs. gerunds, and verbs/clauses/paragraphs as antecedents of event/discourse anaphors.

- (ii) markable boundaries. This applies to titles in Named Entities, genitive-*s*, and fusion of definite determiners with prepositions in German.
- (iii) phrases that are discontinuous or have multiple heads. This occurs in cases of aggregation/summation or conjunctions.

This section is structured as follows: The controversial phenomena listed above will be discussed shortly. Where possible, I will give an estimate of how widespread this phenomenon is. Table 3.2 shows how the different annotation schemes handle these issues.⁷ Line numbers and note numbers in the following text refer to these tables.

Referentiality (line 1 in Table 3.2)

As a general rule, noun phrases (including pronouns) are considered as potentially referring. However, in MUC-7 and OntoNotes, an expression only receives an annotation on the coreference level if it is involved in a linking relation (in OntoNotes, however, the phrases annotated on the syntactic level are available). In the other corpora, all noun phrases in a corpus form markables; it is left to the annotator to exclude the exceptions (e.g. idiomatic expressions, expressions bound by a quantifier).

As to certain kinds of pronouns, there are arguments for and against including them in the set of markables.

Possessive Pronouns (line 2)

On the one hand, possessive pronouns (see Example (96a)⁸) can be used to replace a genitive NP (consider *The instructor's* in (96b) as a replacement of *Her* in (96a).). Thus, they can definitely refer.

- (96) a. Kubeck₁ studied at the Berkeley Hall School in Bel-Air [...]. She₁ began her career as a pilot instructor at small airfields and working at various commuter and freight airlines. Her₁ career breakthrough came in 1989, when she₁ crossed the picket lines at Eastern Airlines.
- b. Kubeck₁ studied at the Berkeley Hall School in Bel-Air [...]. She₁ began her career as a pilot instructor at small airfields and working at various commuter and freight airlines. The instructor's₁ career breakthrough came in 1989, when she₁ crossed the picket lines at Eastern Airlines.

On the other hand, attributive possessive pronouns do not have the status of an NP in most syntactic analyses, like that of the Penn Treebank, TIGER and TüBa-D/Z. See e.g. *Her* in Figure 3.1 (from MUC-7), and *his* in Figure 3.2 (from OntoNotes); the latter figure also shows the analysis of genitive NPs, here, *his daughter's*.

In OntoNotes, the proportion of possessive pronouns in NPs (NPs including possessive pronouns) is 2.29%.

⁷Entries *yes* mean the respective category may form a markable, *no* means it may not form a markable, *impl.* means the scheme gives information only implicitly, e.g. in examples. *n.s.* stands for 'not specified in the respective scheme', *n.a.* for 'not applicable'. Numbers in brackets give page numbers in the original annotation schemes. Footnotes are used for explanations, special conditions, exceptions and examples. Line numbers (first column of tables) and superscript letters are used to organize these additional notes. Markup in examples may be reformatted for uniform appearance.

⁸Example taken from MUC-7. Example (96b) adapted.

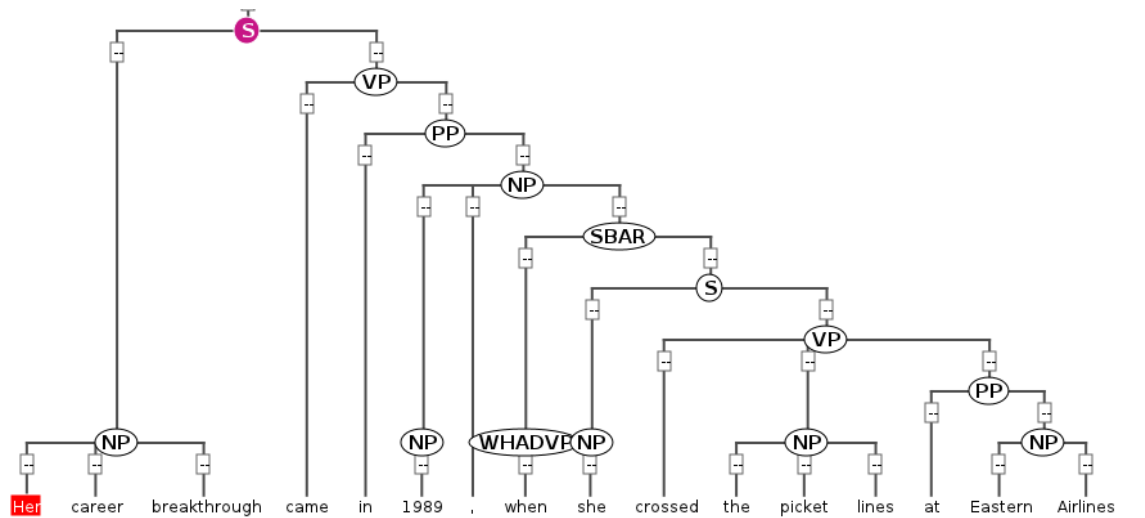


Figure 3.1: Syntactic Analysis of Possessives in MUC-7, visualised in ANNIS2

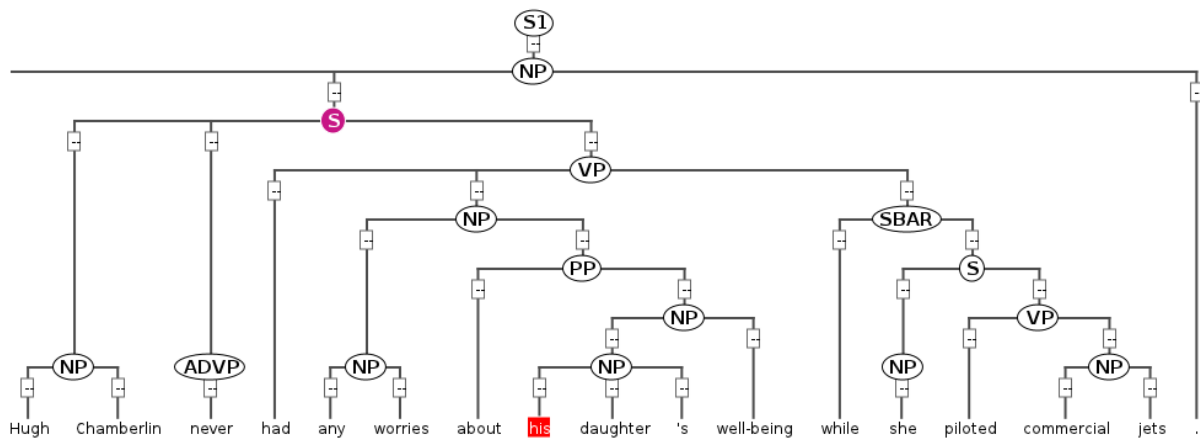


Figure 3.2: Syntactic Analysis of Possessives in OntoNotes 1.0, visualised in ANNIS2

Relative pronouns (line 3)

Relative pronouns function as linkers between a relative clause and the NP being modified or restricted. It is controversial whether relative pronouns form NPs in the syntactic analysis: they do in TüBa-D/Z, but form WHNPs⁹ instead in analyses according to Penn Treebank (which is used in all English corpora presented here). One argument for including them in the set of markables is that relative pronouns in attributive relative clauses are (in the strict sense) coreferent with their antecedents. Also, they can have the

⁹Wh-noun phrases (WHNPs) introduce clauses “with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some wh-word, e.g. *who*, *which book*, *whose daughter*, *none of which*, or *how many leopards*.” (Source: <http://bulba.sdsu.edu/jeanette/thesis/PennTags.html>. Last access April 9th, 2013. Page numbering not available.) Clauses containing the relative pronoun *that* are analysed in the same way. The fact that relative clauses form WHNPs instead of NPs needs to be taken into account during processing. The processing step would then have to distinguish relative clauses from question clauses.

status of an argument to the predicate of the relative clause. As arguments, they and their coreferent descriptions could be of interest, e.g. for the extraction of subcategorization frames of verbs. On the other hand, relative pronouns are syntactically dependent, i.e. they only ever occur *inside* a relative clause (and only the relative clause as a whole specifies the referent). What is more, the coreference relation is not adequate for modeling a linking relation such as (i) syntactic binding, i.e. a linking of the relative clause to the NP it attributes on or restricts (see discussion in Section 2.3.3) or (ii) having the same referential index, i.e. a linking of the relative pronoun to the constituent as a whole (see discussion in Section 2.2.1). In OntoNotes, 1.76% of all NPs have a relative clause with a relative pronoun (an additional 0.77% have a relative clause where the relative pronoun has been dropped). Note that whereas the distinction between attributive and restrictive relative clauses is marked overtly in English (attributive clauses are usually comma-separated), this does not hold for German. Also, relative pronouns cannot be dropped in German.

Reflexive and Reciprocal Pronouns (lines 4 and 5)

Reflexive and reciprocal pronouns have argument status as well. Some reflexive pronouns, however, are required by the verb (see discussion in Section 2.3.3), i.e. they do not represent a referent independently. OntoNotes contains 1.00% reflexive and 0.01% reciprocal pronouns (of all NPs).

Zero Forms/Traces (line 6)

According to some syntactic analyses (including the Penn Treebank analysis), traces are used to represent arguments that have moved (see discussion in Section 2.3.3). Zeros are used to represent arguments not realized in the utterance. I do not consider traces or zero forms referents. In OntoNotes, 4.65% of all NPs are traces (labeled *T*), and 5% are zeros (labelled *PRO*).

Interrogative Constituents (line 7)

Like the pronouns listed above, interrogative constituents can also have argument status; however, it is debatable whether they refer (see also Section 2.2.6). The proportion of interrogative NPs is 0.31% in OntoNotes.

Temporal, Locational and Numeric Expressions (lines 8 to 10)

Expressions referring to time and place can be realized as NPs, e.g. *yesterday*, *next week*, *last year* or *home*, *Cambridge, Mass.*, *editorial page*. Alternatively, they can be realized as adverbial phrases (ADVPs), such as *here*, *there*, *abroad*, *nearby*, *below*, as well as *now*, *then*, *currently*, *recently*, *early*, *late*, *so far*, *two years ago*, *once*. Usually, ADVPs are not considered as referring. According to the Penn Treebank scheme, however, adverbs like *tomorrow* and *yesterday* are part of speech tagged *NN* (common noun). Their German counterparts *morgen* and *gestern* are tagged *ADV* (adverb). Equivalent entities may therefore be assigned different categories depending on the tag set that is used. An estimation is hardly possible based on the given annotation.

Numeric expressions, like quantified expressions, raise the question whether they refer specifically (see discussion in Section 2.2.5). In OntoNotes, 2.86% of all NPs have a cardinal number as a modifier.

Embedding of Expressions (line 11)

Referring expressions may embed other referring expressions (see discussion in Section 2.2.1). However, there is controversy as to whether they should be marked for coreference, and which elements (e.g. modifying NPs/PPs, premodifiers of compound

nouns; names as part of complex names) should be marked. In OntoNotes, 39.59% of all NPs are complex (29.04% of all NPs embed NPs; 10.55% embed PPs). 1.77% of coreferent expressions (expressions with a coreferent mention in the left or right context) embed other coreferent expressions. (Note, however, that these numbers do not include premodifiers or names embedded in other names, as they are not annotated in OntoNotes as a rule, see note 11b. in Table 3.2).

Expressions Derived from Verbs; Clauses and Paragraphs (lines 12 to 15)

The potential to refer is usually attributed to nominal word forms (see Section 2.2.1). It is controversial whether present participles and gerunds refer. The same applies to verbs, clauses, and paragraphs¹⁰. They can describe more or less complex events, states or situations that can, as a whole, be referred to again. Quantifying the proportion of non-nominal, potentially referring, elements, on the basis of existing annotation is possible only for verbs. In OntoNotes, 6.55% of all first mentions are verbs (in 2.47% of all identity relations, the antecedent is a verb).

Markable Boundaries

There is some disagreement regarding which parts of a mention a markable should include.

Titles (line 16)

Titles before names do not form part of the NP to be annotated in MUC-7 (note 16a. in Table 3.2). In OntoNotes, however, they do: there, 0.25% of all NPs are names that have a common noun modifier.¹¹

Possessive 's or ' (line 17)

Possessive NPs in English end in an 's or ' (the latter representing the plural form, or singular form for words ending in *s* or *x*). According to some tokenization conventions (including those used in the Penn Treebank), these suffixes represent separate tokens. These tokens form part of the same NP (and thus form part of the markable if NP boundaries from syntax are re-used for coreference) in OntoNotes, but not in MUC-7. In OntoNotes, 2.25% of all NPs are possessive. In German, apostrophes are used only for forms ending in *s* or *x*. Tokenization usually does not separate such tokens, the ending forms a regular part of the token it is attached to.

Fusion (line 18)

German has fusion of prepositions and definite determiners. Fusion is not always optional, even in written language. The TüBa-D/Z annotation scheme defines a noun phrase consisting of the noun only, excluding the definiteness feature. This noun phrase forms a prepositional phrase together with the fused preposition-determiner complex (see Figure 3.3). 4.12% of NPs are preceded by a preposition-determiner complex.

Discontinuous or Multiple-head Phrases (lines 19 and 20)

The mapping between phrases and referents is not always 1:1. One referent may be represented by several (parts of) phrases (combined to so-called discontinuous phrases), and one phrase may refer to more than one (group of) referent(s). For discontinuous constituents, Chiarcos and Krasavina (2005) give the following examples:

(97) You'll meet a man tomorrow carrying a heavy parcel

¹⁰Here, I use the term *paragraph* for a sequence of sentences or clauses; these need not necessarily form a typographic paragraph.

¹¹Note that in OntoNotes, titles like *Mr.*, *Rep.*, *Sen.* as well as *Senator* etc. are tagged *NNP* as part of the named entity. These names which include titles amount to an estimated additional 1.67% of all NPs. This number is estimated using heuristics to search for the open class of titles.

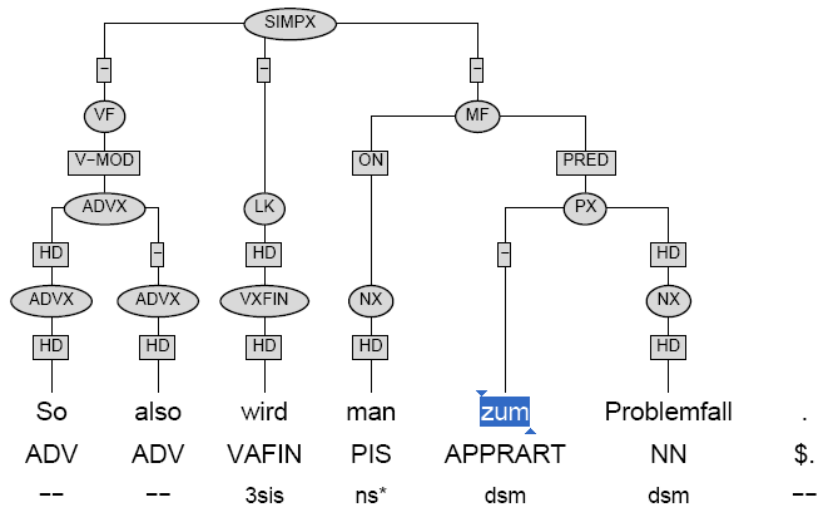


Figure 3.3: Syntactic analysis of a PP with Fusion of Preposition and Definite Determiner in TüBa-D/Z (adapted from Telljohann et al. 2009, page 29). ‘So this is how one becomes a problematic case.’

(98) *Bücher hat Anna drei.*
 books has Anna three

Anna has three books./As for books, Anna has three of them. (split-NP)

As to multiple-head phrases, suppose one phrase is used to refer to more than one (groups of) referent(s), e.g. the conjunction *young men and women*. Then it is possible to refer back to one (group of) referent(s) separately later on in the text, e.g. *the young women*. This group of referents has been mentioned before, but the expression is structured in a way that does not make it trivial to form a markable.

In ARRAU, it is possible to create a markable from parts of expressions (see note 19b. in Table 3.2).

To resume this section, each of these phenomena may seem marginal when considered in isolation. Most of the proportions range around 2%. In sum, however, they lead to considerable disagreement between the different resources (cf. Table 3.2).

0.	MUC-7 (Hirschman and Chinchor 1997)	OntoNotes (Weischedel et al., 2007)	Nissim (Nissim, 2003)	ARRAU ^a
	candidate markables	yes	yes	yes
1.	preconditions	no	no	no
2.	possessive pronoun	yes (p. 4, 6)	yes (p. 1)	yes (1c)
3.	relative pronoun	no (p. 7) ^a	yes (p. 4)	yes (1a)
4.	reflexive pronoun	yes (p. 6) ^a	yes (p. 5)	n.s. (probably yes)
5.	reciprocal pronoun	n.s.	n.s.	n.s. (probably yes)
6.	zero forms/traces	no (p. 7)	no (p. 1)	no (1a)
7.	interrogative constituent	no (p. 5)	yes (p. 1, 11)	no (3w)
8.	temporal expression	only NPs (p. 5)	unclear (p. 1) ^b	n.s. (probably yes)
9.	locational expression	only NPs (p. 8)	unclear (p. 1) ^b	yes (4f)
10.	numeric expression	yes (p. 4, 5)	yes (p. 7)	no ^a
11.	embedded expression	partly (p. 5f., 9) ^a	n.s. (no, p. 8)	partly ^c
12.	present participle	yes (p. 6) ^a	no (→ verbs)	no (p. 5)
13.	gerund	no (p. 6) ^a	no (→ verbs)	no (p. 5)
14.	verb	no (p. 2, 5)	no (impl., p. 5)	no (p. 5)
15.	clause/paragraph	no (p. 2, 5)	no (impl., p. 5)	only as antecedent (p. 12)
16.	titles in Named Entities	no (p. 8) ^a	n.s.	n.s.
17.	possessive <i>s</i>	no (impl., p. 6, 13) ^a	n.a.	n.s.
18.	fusion ^a	n.a.	n.a.	n.a.
19.	discontinuous markable	no (p. 9) ^a	n.s.	in conjunctions (1e) ^b
20.	conjunction	yes (p. 7) ^a	yes (p. 7)	no, separately ^b

Table 3.2: Markable Definitions (part I: English Corpora)

^{0a}. Sources: for `_ARRAU_manual.txt` (their example numbers); `man_anno_1_trains.pdf` (page numbers)

^{1a}. Reference counts as specific if any of the mentions in the coreferential chain is specific. Specificity is determined by the respective expression's form: specific expressions include names and syntactically definite descriptions as well as pronouns (with the exception of generic *you*; p. 3). Verbs do not count as specific (p. 4); they are thus only annotated if taken up by a specific expression.

^{1b}. Temporal expressions are excluded (see note 8b). Idioms and expletives receive the label *non-applicable* (p. 3).

^{1c}. The annotation process starts from nominal anaphora (p. 1).

^{3a}. The maximal NP forms the markable, which includes non-restrictive relative clauses among others.

- 4a. Including emphatics like in He₁ *is*, himself₁, unsure of the outcome.
- 4b. “all pronouns” are markable.
- 6a. Traces are only contained in the parser output, not in the files for coreference annotation.
- 8a. “Temporal expressions are eligible for co-reference, including deictic expressions such as: *now, then, today, tomorrow, yesterday*, etc. or other temporal expressions that are relative to the time of the writing of the article.”
- 8b. Nissim excludes “those [NPs] subcategorized as ADV, LOC, TMP, or DIR, and those within ADVPs subcategorized as LOC, DIR, TMP” (p. 1). However, *tomorrow, Monday, every day, all the time* (p. 6, 10) in her examples are annotated.
- 9a. with the exception of embedded location names, see *embedded expression*, line 15 in this table.
- 9b. cf. note 8b; however, *at home, at work* (p. 6) are annotated.
- 10a. Numeric expressions are considered indefinite (3d) and can thus only serve as antecedents.
- 11a. Including premodifiers, e.g. *He was accused of money laundering and drug trafficking. However, the trade in drugs... and parts of words London ... London-based ...* Expressions embedded in named entities (which include date expressions) are not considered markables for reasons of consistency with the named entity (NE) task. Negative examples (marked by dashed underlines): *Equitable of Iowa*₁; *In a report issued January 5, 1995*₂, *the program manager said that there would be no new funds this year*.
- 11b. Only embedded NPs (no premodifiers of compounds etc.); “Proper names are considered to be atomic, and nested mentions inside proper names are not annotated separately. In the following examples, the location names that form part of the organization names are not eligible for co-reference [...] *Massachusetts Institute of Technology, Bank of America, the Chicago Board of Trade*” (dashed lines added to mark negative examples). Also, “[a]djectival premodifiers are not marked. [...]”: *U.S. in U.S. economy* is not a proper NP; it can be replaced by *American*, thus has an adjectival function.
- 11c. In the case of identical referents, the maximal phrase is annotated. Locations of companies (5.), possessives (1c, see above), and specific premodifiers (1d) form markables of their own.
- 12a. “Present participles which are modified by other nouns or adjectives (*program trading, excessive spending*), are preceded by an article (*a, the, my, etc.*) or are followed by an ‘of’ phrase (*slowing of the economy*) are to be considered noun-like and ARE markable.”
- 13a. “A phrase headed by a present participle is taken to be verbal if it can take an object [...] or can be modified by an adverb.”
- 14a. Verbal elements are only annotated for coreference if the referent is mentioned as an NP at least once in the text; “Only the single-word head of the verb phrase is included in the span, even in cases where the entire verb phrase is the logical co-referent.”
- 16a. If the head is a name, the entire name is marked, incl. suffixes such as *Sr., III*, etc. on personal and *Corp.* on organization names, but not titles or modifiers.
- 16b. Titles form part of the NP and are thus automatically included in the markables, e.g. *Mr. Smith*.
- 17a. In *the man’s arm*, there are two markables, *the man* and *the man’s arm*. Another example is *The White House*₁ *sent its health care proposal to Congress yesterday*. *Senator Dole said the administration*₁ *’s bill had little chance of passing*.
- 17b. “Noun phrases extracted from the treebank may include the possessive ‘s in the NP. The ‘s ending should be included in the extent of the noun phrases that are co-referenced”, e.g. *Iowa*_{’s} *governor ... Postville, Iowa*
- 18a. Phenomenon only applies to German, see part II of this table.
- 19a. Discontinuous phrases (*president of Amalgamated Text Processing Inc.*) may be split up: *Ms. Fribble*₁ *was president*₁ *and CEO of Amalgamated Text Processing Inc.*₁.
- 19b. “(Tentative) use discontinuous markables for noun or premodifier coordinations”, e.g. *the Singapore and Kuala Lumpur stock exchanges, or the Singapore and Kuala Lumpur stock exchanges*, respectively.
- 20a. conjoint or individual markable depending on coreference (see note 19a above).
- 20b. each element is annotated separately; markables can be conjoined to sets.

	PCC (Krasavina and Chiarcos, 2007) ^a	TüBa-D/Z (Naumann, 2006)	DIRNDL (Riester et al. 2007, Riester et al. 2010) ^b
0.	yes	yes	
candidate markables	primarily definiteness ^a	definiteness ^b	primarily definiteness ^c
1. preconditions	yes	yes	
2. possessive pronoun	yes (p. 26)	yes (p. 3, 7)	yes (p. 14)
3. relative pronoun	partly (p. 28) ^a	yes (p. 3)	yes (Fig. 4)
4. reflexive pronoun	partly (p. 26) ^a	yes (p. 3) ^b	yes (p. 7)
5. reciprocal pronoun	n.s.	yes (p. 3)	yes (p. 7)
6. zero form/trace	no (not in grammar)	no (p. 2)	no (not in grammar)
7. interrogative constituent	n.s.	no (p. 18) ^a	n.s.
8. temporal expression	n.s.	only NPs (<i>impl.</i> , p. 3) ^a	yes (p. 8)
9. locational expression	yes (<i>impl.</i> , p. 28) ^a	only NPs (<i>impl.</i> , p. 15)	yes (p. 3, 5, 7)
10. numeric expression	n.s.	yes (<i>impl.</i> , p. 19)	yes (p. 10) and also adjectives
11. embedded expression	yes (<i>impl.</i> , p. 27f.)	yes (p. 4)	yes (p. 3) ^a
12. present participle ^a	n.a.	n.a.	n.a.
13. gerund ^a	n.a.	n.a.	n.a.
14. verb	no (p. 25)	no (p. 2, 8)	yes (p. 12)
15. clause/paragraph	no (<i>impl.</i> , p. 39)	no (p. 3)	yes (p. 12)
16. titles in Named Entities	yes (p. 27)	yes (p. 3)	yes (p. 5, 6, 13)
17. possessive <i>s</i> ^a	yes (p. 27)	yes (p. 4)	yes (p. 3) ^b
18. fusion	yes (p. 28) ^a	yes ^b	yes (p. 3, 8) ^c
19. discontinuous markable	yes (p. 29)	n.s.	n.s.
20. conjunction	yes (p. 28)	yes (p. 3)	yes (p. 12) ^a

Table 3.2: Markable Definitions (part II: German Corpora)

- 0a. (updated January, 2013)
- 0b. Page numbers refer to the earlier version of the guidelines unless stated otherwise.
- 1a. Definite NPs form *primary markables*, antecedents that are not definite form *secondary markables*.
- 1b. “definite NPs, including complex (e.g. coordinated) noun phrases” (p. 3)
- 1c. (p. 4f.) In recent versions, the relation *indef-rel* is included.
- 3a. Only in possessive constructions: *Und so schielten die Israelis nach Washington, an dessen Tropf sie hängen.* - ‘And so the Israelis were looking towards Washington, whose drip-feed they are on.’
- 4a. unless the reflexive is required by the verb (test diagnostic: required reflexives cannot be topicalised).
- 4b. exceptions: emphatics (p. 18) and “cases of ‘inherently reflexive verbs’, as e.g. ‘sich ereignen’ (eng. ‘happen (itself)’)” (p. 17)
- 7a. “we only annotate those cases [of interrogative pronouns] where these elements function as a relative pronoun, as e.g. in: ‘Ich interessiere mich für das₁, was₂ du sagst.’ - eng.: ‘I am interested in that₁ which₂ you are saying.’ In this case, markable 2 stands in a ‘bound’ relation to markable 1.”
- 8a. e.g. *last winter, then.*
- 9a. e.g. *Washington*, see example in note 3a.
- 11a. If parts of a phrase refer to a different object than the phrase as a whole, they should receive a label of their own. *Sie werden von der regulären Armee Kroatiens unterstützt.* - ‘They are supported by Croatia’s regular army.’
- 12a. Phenomenon only applies to English. Nominalization is marked by capitalization in German, which allows a clear distinction of nominal vs. verbal or adjectival uses of words.
- 13a. Phenomenon only applies to English. See note 12a.
- 17a. in German, the *s* is not preceded by an apostrophe.
- 17b. *Kroatiens* - ‘Croatia’s’
- 18a. As a general rule, prepositions are included in the markable where present.
- 18b. though not explicitly mentioned in the guidelines, in the corpus, prepositions bearing a definiteness markers are systematically included in the markable (in contrast to “im Metropol”, p. 12)
- 18c. including pronominal adverbs if they refer to a relative clause in the following text (*Die Parteiverantwortlichen sollten darüber reden, wie man aus diesen Interview-Kriegen herauskommt.* ‘The party leadership should talk about how to get out of those interview wars.’), p. 7
- 20a. If all phrases have the same information status, the phrases as a whole is also labeled with this information status.

3.2.1.2 Preconditions

An explicit marking of relevant entities (entities meeting the preconditions) has two advantages:

- (i) during the annotation process, completeness of the annotation can be checked: every markable must receive an annotation, a markable left unannotated signals it might have been overlooked.
- (ii) using the corpus, one can distinguish expressions that are non-referring (or non-specific or indefinite, respectively) from expressions whose referents are mentioned just once (so-called ‘singletons’).

Commonly, referentiality is considered a necessary precondition for coreference across all annotation schemes. In MUC-7 and ARRAU, referentiality is the only precondition.¹² Nissim (2003) additionally excludes adverbial, temporal, locational and directional NPs (see note 8b. in Table 3.2). OntoNotes and TüBa-D/Z require specificity/definiteness. They do not point to the literature for definitions of specificity/definiteness, but give the following definitions: in OntoNotes, specificity is defined depending on the surface form. Proper nouns, referring pronouns and demonstratives, NPs with a definite determiner, and indefinite specific NPs (e.g. *a man I know*) are considered specific. “Bare plurals [...] are always generic” (Authorless, 2007, p. 4). Annotators are instructed to annotate an expression if its referent is mentioned at least once using such a specific expression. Earlier (or later) mentions may be bare plurals, indefinite singulars, or verbs. In TüBa-D/Z, in contrast, only “definite descriptions, i.e. definite NPs, including complex (e.g. coordinated) noun phrases” (p. 3) are linked to their antecedents. Expletives are labeled as such. In PCC, non-referring expressions are labeled as such. In DIRNDL, this is also the case (with the exception of expletives, which are annotated EXPLETIVE). Definite expressions are differentiated from indefinite expressions (indefinites receive a label with the prefix INDEF).

Neither OntoNotes nor MUC nor TüBa-D/Z make the annotator’s decision on an expression’s referentiality explicit (with the exception of expletives in TüBa-D/Z). Entities excluded from the annotation per definition of the scheme and singletons are equally left unannotated. The original reason for an NP left unannotated cannot be reconstructed. In Nissim’s scheme, in contrast, markables where the text material is unclear are labeled *unclear*, and non-referential markables are labeled *not applicable*. In ARRAU, non-referring expressions are also annotated. They receive the label *non-referring*, along with their subtype: *expletive*, *predicate*, *quantifier*, or *idiom*. In the PCC coreference annotation, only nominal reference is annotated. Markables are organized in two groups: referential definite NPs constitute *primary markables*; antecedents that are not definite NPs constitute *secondary markables*. VPs do not form markables. Expletives and parts of idiomatic or collocational expressions are left unannotated. Parts of productive metaphors, however, may form markables.

¹²In reliability studies on ARRAU, temporal expressions have been excluded (Poesio and Artstein, 2005).

3.2.2 Coreference and Context

Coreference of expressions means that these expressions refer to the same thing, i.e. identity holds between their respective referents (for a more formal definition and discussion, see Section 2.3). Coreference, as well as discourse-givness, depends the notion of context, i.e. which parts of this context are considered as antecedent candidates.

3.2.2.1 Discourse-Givness and Coreference

An overview of the definitions of discourse-givness, information status and coreference is given in Table 3.3 (see Table 2.3 for the underlying concepts and definitions).

In MUC-7, all forms of identity between nominal elements are annotated, including coreference between kinds or abstract concepts, binding, predications, asserted identity, appositions and function-value relations. The relation `IDENT` (short for *identity*) is used for all of these cases. In OntoNotes, the `IDENT` relation is only used if the referent has been mentioned as a specific expression at least once (neither between abstract entities, nor between generic *you* or generic indefinite NPs). An antecedent may be a verb. For appositions, the `APPOS` relation is used.

According to Nissim’s (2003) scheme, only specifically referring and kind-referring anaphoric NPs are linked to their antecedents. One should “not annotate coreference links for ‘I’ and ‘you’ and their forms” (Nissim, 2003, p. 3). Neither should links to VPs or predicative phrases be created. As to expressions related to the context via bridging, binding or the function-value relation, they are annotated as *mediated* or *func_value*, respectively, but not linked to their antecedents.

In ARRAU, coreference between specific expressions, kinds, abstract concepts, and events, as well as proposition anaphors, binding relations, and even bridging is annotated. The anaphor is labeled *old*. Expressions related to the context via bridging or not related at all are labeled *new*. Genericity is also annotated

In TüBa-D/Z, several relations are distinguished: *anaphoric* for the relation of definite anaphoric pronouns to their antecedents, *cataphoric* between a cataphor and its first full mention, *bound* for the relation of a bound pronoun to its binder, *coreferential* for coreference between NPs where one of these mentions is specific. The relations *instance* and *split_antecedent* represent bridging relations.

PCC’s coreference annotation only extends to nominal coreference, i.e. links are created only between elements that are noun phrases. Coreference is defined via the substitution test: for expressions to be coreferent, they need to be substitutable by one another.

DIRNDL’s focus is on coreference between definite expressions. The annotation scheme, however, has been adapted to include indefinite expressions as well. Bridging relations are also annotated, referring expressions not related to the context are labeled *new*.

How the linking is realized in the different corpora will be discussed in Section 3.2.3.

	MUC-7 and Clinchor, 1997		Hirschman (Authorless, 2007)		Nissim (Nissim, 2003)		ARRAU documentation ^a	
	link	relation label	link	relation label	link	relation label	link	relation label
1. non-referring ^a	no	no	no	no	yes (p. 2) ^b	no	no	
2. cataphor	+	IDENT <i>impl.</i> (p. 6) ^a	n.s.	n.s.	n.s.	n.s.	n.s.	
3. binding	+	IDENT (p. 10)	+	IDENT (p. 3, 9) ^a	+/-	old or med ^b	+	old (p. 8) ^c
4. event anaphor	-	partly no (p. 2, 3, 5) ^a	+	IDENT ^b	-	old/event ^c	+	old (p. 8) ^d
5. proposition anaphor	-	no (\rightarrow events)	-	no (\rightarrow VPs)	-	n.s. ^a	+	old (p. 12)
6. predication/asserted identity	+	IDENT (p. 11) ^a	-	no (p. 6) ^b	-	no (p. 10)	-	no (p. 6)
7. function-value	+	IDENT (p. 13) ^a	-	no ^b	-	med/func_value (p. 7)	-	no (n.s.)
8. apposition	+	IDENT (p. 11) ^a	+	APPOS (p. 7) ^b	-	no (p. 6)	-	no ^c
9. between kinds	+	IDENT (p. 12f.)	+	IDENT partly ^a	+	old/ident-generic ^b	+	old ^c
10. between abstract concepts	+	IDENT ^a	-	no (p. 2, 3)	-	n.s.	+	old (p. 8)
11. between spec/def. NPs	+	IDENT ^a	+	IDENT ^b	+	old ^c	+	old (p. 8)
12. bridging	-	no	-	no	-	med ^a	+	old; new (p. 10) ^b
13. no relation	no	no	no	no	new (p. 10) ^a	new (p. 8)	new (p. 8)	

Table 3.3: Definitions of Coreference (part I: English Corpora)

^{0a}. sources: for_ARRAU_manual.txt (their example numbers); man_anno_1_trains.pdf (page numbers)

^{1a}. Are non-referring expressions explicitly marked up?

^{1b}. Nonreferring NPs, i.e. pleonastic “it” (dummy-it, or ‘it’ in extrapositions or clefts) and pronouns and nominals in idiomatic phrases (“in [fact], [there] is, [...] by [accident], [...] [you] know”) are marked ‘not applicable’. Note that adverbial, temporal, local and directional phrases are excluded from the annotation (see note 8b. in Table 3.2).

^{2a}. “*There is no business reason for my₁ departure*”, *he₁* added.

^{3a}. Binding phrases open up “distinct Ident chains, each containing a generic and the referring pronouns.” (p. 4)

^{3b}. Expressions bound in generic utterances are annotated as ‘old/ident-generic’ and linked accordingly. Expressions bound by quantifiers are annotated as med/bound (no link). Relative pronouns are marked ‘old/relative’ and linked accordingly.

^{3c}. link to antecedent (if quantified, the domain of quantification).

^{4a}. unless nominal or gerund (p. 6)

^{4b}. “This includes morphologically related nominalizations [...] and noun phrases that refer to the same event but are lexically distinct [...]. Only the single-word head of the verb phrase is included in the span, even in cases where the entire verb phrase is the logical co-referent.” (p. 2)

- 4c. “A: I like shopping on line. B: Yeah - I like it too.” (p. 5)
- 4d. “Examples of abstract objects are facts, events, actions, and plans. For instance, in the example [...] [*M: I want you to take [a boxcar] from Elmira and load [it] with oranges*], M is proposing a plan: to take a boxcar from Elmira and loading it with oranges.” (p. 8)
- 5a. not in extraposition “It’s good that you cleaned up” (p. 2)
- 6a. “if the text asserts them to be coreferential at ANY TIME”
- 6b. “In a copular structure, the subject is added to the co-reference chain [if further mentions follow].” (p. 6)
- 7a. When more than one value is specified, the annotator should mark “the most ‘current’ value in its clause” (p. 13)
- 7b. not in asserting sentences, only in cases of ‘proper coreference’, incl. premodifiers (“*The company’s [\\$150]_x offer was unexpected. The firm balked at [the price]_x.*”), appositions and “appositive-like mentions” (“*The price]_x? [\\$300 *U*]_x. [A lot]_x by current standards.*”) (p. 11)
- 8a. “only when they constitute a separate noun phrase following the head.” (p. 11)
- 8b. The maximal phrase is linked to its head; this head can enter IDENT relations. The head is determined by means of a “specificity scale” (“Proper noun>Pronoun>Def. NP>Indef. spec. NP>Non-spec. NP”, p. 7): the most specific mention (if equivalent, the left-most mention) is the head. Multiple appositives can be attributed to one head. Appositions include numeric ages, relative clauses, and noun phrases.
- 8c. “the embedding NP would be chosen as an antecedent of subsequent anaphoric references, rather than the NP in appositive position.” (Poesio, 2004c, p. 5).
- 9a. Only “a generic and the referring pronouns” (p. 4) are linked; neither full nominal generics nor generic ‘you’ (p. 3) are linked.
- 9b. “[i]f a generic is coreferential with a generic that has already been introduced” (p. 5).
- 9c. “A term-denoting markable in the sense discussed above, i.e., which refers to a concrete or abstract object which has already been mentioned or discussed earlier in the dialogue.” (p. 8)
- 10a. “program trading”, “excessive spending” (p. 6), “loss” (p. 16)
- 11a. criterion: they “refer to the same object, set, activity, etc.” (p. 10)
- 11b. exception: adjectival premodifier (*U.S. economy=American economy*, p. 2). “Prenominal modifiers (e.g., [...] ‘the [ocean drilling] company’) are markable only if either the prenominal modifier is coreferential with a named entity or to the syntactic head of a maximal noun phrase. That is, there must be one element in the coreference chain that is a head or a name, not a modifier.” (p. 6)
- 11c. Entities are annotated with their information status. (Discourse-)old expressions include entities mentioned previously in the discourse (which are additionally marked ‘ident’), as well as “the personal pronouns ‘I’ and ‘you’ in their referential usage” (additionally marked ‘general’) and generic uses of pronouns (p. 4). ‘Old’ entities are linked to their antecedents, except “I’ and ‘you’ and their forms” (p. 2).
- 12a. med/event for possible roles associated to an event previously mentioned (“We were *travelling around Yucatan*, and **the bus** was really full.”, p. 7), med/aggregation for coordinations of entities where at least one is old, but not all (p. 7), med/set for all kinds of set relations, including hyponyms and hypernyms, using WordNet as a reference (p. 7), med/poss for entities that are in a possession relation to another entity mentioned (p. 8), med/part-whole for physical objects that form parts of entities previously mentioned (meronymy in WordNet as a reference, p. 8/9), and med/situation for entities that form part of the “situation set up by the antecedent” or for abstract objects that are part of other entities mentioned (p. 9), “at *the wedding, the bride* ...”, WordNet definitions and FrameNet are used as a reference.
- 12b. *Old* for summation of referents, e.g. “*Please hook up the engine and the boxcar and send them old to Elmira.*” (p. 14), with links to multiple phrases. *New* for referents related to the context by bridging relations like *part-of*, *set relations* (elements or subsets of sets), *other* (containing the word *other*), and *misc* for cases to be discussed.
- 13a. exceptions: old/generic (p. 4f.) med/general for “generally known entities such as ‘the universe’, ‘the moon’, ‘the people’ (if not used specifically) and similar. Time-related expressions such as ‘tomorrow’, ‘Monday’ and so also fall in this category” (p. 6).

	PCC		TüBa-D/Z		DIRNDL	
	link	phrase label	link	relation label	link	phrase label
0.		(Krasavina and Chiarcos, 2007)		(Naumann, 2006)		(Riester et al. 2007, 2010) ^a
1. non-referring ^a	yes (p. 34)		partly ^b		partly ^c	
2. cataphor	+	(p. 32) ^a	+	cataphoric ^b	+	d-given (p. 7) ^c
3. binding	-	no (<i>impl</i> , p. 31) ^a	+	bound ^b	+	d-given(-pronoun) ^c
4. event anaphor	-	no (p. 25)	-	no (p. 8)	+	d-given (-pronoun) (p. 11)
5. proposition anaphor	-	referentiality: referring (p. 31)	-	no (p. 8)	+	d-given (-pronoun) (p. 11)
6. predication/asserted identity	-	referentiality: other (p. 34)	+	coreferential ^a	-	new ^b
7. function-value	n.s.		-	no (p. 5, 6)	-	no (p. 15) ^a
8. apposition	-	no (p. 26) ^a	-	no (p. 12) ^b	?	yes (p. 13) ^c
9. between kinds	-	referentiality: other (p. 34)	-	no (p. 13)	+	indef-rel (p. 10) ^a
10. between abstract concepts	?	only if definite	+	coreferential (p. 13)	+	indef-rel (p. 10)
11. between spec/def. NPs	+		+	coreferential/ anaphoric (p. 12, 17) ^a	+	D-GIVEN ^b
12. bridging	-		+	partly (instance/ split_antecedent) ^a	+	BRIDGING ^b
13. no relation	no		no			NEW/ACCESSIBLE/ SITUATIVE/INDEF

Table 3.3: Definitions of Coreference (part II: German Corpora)

^{0a}. Page numbers refer to the earlier version of the guidelines unless stated otherwise.

^{1a}. Are non-referring expressions explicitly marked up?

^{1b}. Instances of non-referring ‘*es*’ (*‘it’*), e.g. subjects of weather verbs and presentational constructions, are marked ‘expletive’ (p. 35).

^{1c}. expletives: EXPLETIVE; *nothing*, *nobody*; NULL.

^{2a}. pointing from left to right.

^{2b}. “A cataphoric relation holds between a pronoun referring to a following antecedent within the same or the following sentence.” (p. 21)

^{2c}. In this case, the ‘antecedent’ is to the right of the target.

^{3a}. Instances of binding fail the substitution test.

^{3b}. Expressions bound by a quantifier (p. 6) or generic *man* (*‘one’*, p. 23) are annotated as ‘bound’, other generic expressions are not (p. 24).

^{3c}. (n.s. for quantifiers; kind-referring: antecedent accessible-general-type, anaphor d-given(-pronoun) (p. 8); RELATIVE for relative pronouns in non-restrictive relative clauses)

-
- 6a. only if predicative NP is (syntactically) definite (p. 13); exception: constructions with *as* (p. 13)
- 6b. (If the relation between two expressions is established only by the sentence construction (e.g. copula), this relation is not annotated, as the second expression itself represents new information (p. 15): *Der EU-Kommissar bezeichnete Deutschland als treibende Kraft in Europa.* ‘The EU Commissioner called Germany a driving force in Europe.’)
- 7a. *Die Kosten belaufen sich auf 4 Millionen Euro.* (‘The costs amount to 4 million Euros.’)
- 8a. Only the maximal phrase is annotated.
- 8b. “Appositions belonging to the same maximal NP [receive] no internal reference marking”
- 8c. Each maximal phrase with a referent of its own is labelled.
- 9a. not in the case of copula (p. 15, *Katzen sind Säugetiere.* ‘Cats are mammals’.)
- 11a. anaphoric for definite pronouns referring back to a contextual antecedent, coreferential for other NPs. Exception: mention is embedded in a named entity (p. 15).
- 11b. (-PRONOUN,-REFLEXIVE,-SHORT,-REPEATED,-EPITHET, p. 5-7; synonymous removed, epithet and reflexive added in (Riester et al., 2010), p. 718, if both labels -pronoun and -repeated are applicable, -pronoun is to be chosen.)
- 12a. *instance* “where a specific pronoun or NP refers to a particular instantiation of the class identified by an NP” (p. 31), *split-antecedent* for relations “between coordinate NPs (e.g. ‘Jane and Mary’) or plural pronouns (e.g. ‘both’) and pronouns/definite NPs referring to one member of the plural expression.” (p. 29), other bridging relations are not annotated (p. 5)
- 12b. (-0, -TEXT, -CONTAINED, Riester et al. (2010) p. 718)

3.2.2.2 Notions of Context

The annotation schemes take different kinds of context into account. As a result, some expression α in some text c may receive different labels (and/or links) according to different annotation schemes. In MUC-7 and ARRAU, an NP’s antecedent is to be sought among all nominals, including premodifiers, in the (left) co-text¹³ of this NP. In ARRAU, besides nominals, propositions are considered. In OntoNotes, all noun phrases and verbs in the co-text are regarded; in PCC’s coreference scheme, all noun phrases are considered. Nissim’s annotation scheme defines relations from NPs to the co-text and consequences thereof (from propositions, events, aggregations, through to frames). Additionally, it uses the non-textual context, i.e. the utterance situation, and world knowledge (e.g. ‘the universe’ and proper names referring to generally known entities). The TüBa-D/Z scheme allows for relations to nominals in the co-text, as well as some consequences (like instantiation (*‘instance’*) and aggregation (*‘split_antecedent’*), and the right co-text in case of cataphors. Similarly, in DIRNDL and PCC’s information status annotation, relations to noun phrases in the co-text are considered, including bridging relations (e.g. the subset relation etc.).

Using only the co-text is the easiest operationalizable option; taking consequences into account needs extra criteria, e.g. WordNet or FrameNet relations (Nissim, 2003), or topic maps (Goecke et al., 2007) to ensure consistency.

3.2.3 Formalization and Evaluation of the Annotation

Trivial as it may seem, the interpretation and exploitation of a corpus depends on what is annotated and how it is annotated, as well as how consistent the annotation is, and how well it is documented.

3.2.3.1 Formalization

Annotation schemes differ as to whether they demand an explicit annotation of the relation itself, i.e. a linking of entities. The structural model of annotation has consequences on the interpretation of the annotated data, both on contentual and on technical side.

Contentual aspects

An explicit linking to the antecedent or binding expression has a practical advantage: it allows for a certain entity that its related entities (typically the antecedents) be retrieved. As a consequence, the properties of these related entities are accessible, e.g. whether an entity is NP (as opposed to a VP or S), a pronoun, definite, quantified, etc. This allows for an automatic assignment of labels, which facilitates the comparison to categorizations in other schemes. For instance, event anaphors are allowed in OntoNotes but not in MUC-7; excluding event anaphors could be one step towards a better comparability between the two resources.

Technical aspects

Two different representations of coreference have been used in the corpora presented

¹³By co-text, I mean the text within the same document. (This term is introduced to distinguish textual and situative context.) All schemes take the whole document into account, rather than a window of text.

above: (i) a coreference set: each referent¹⁴ has an ID. Every mention of a referent is annotated with this referent’s ID. (ii) a directed graph: each expression has an ID. A coreferent expression is annotated with its antecedent’s ID.¹⁵ In any case, the edge type can be specified.

Coreference sets (option i) can be interpreted as representing equivalence classes, implying symmetry and transitivity. This model is adequate for strict coreference, but not adequate for binding relations and 1:n relations (multiple antecedents). It is used in OntoNotes in each of the relations IDENT and APPOS.

A directed graph (option ii, the so-called chain-model) is used in MUC-7, Nissim (2003), ARRAU, TüBa-D/Z and PCC’s coreference annotation.

In contrast to these schemes, PCC’s information status annotation (Götze et al., 2007) demands ‘flat’ annotations, i.e. labels (attribute/value pairs), instead of pointing structures.

For the purpose of this work, any annotation will automatically be turned into such ‘flat’ annotation labels during preprocessing. These labels specify (i) whether or not an entity has been previously mentioned (i.e. has a co-referring expression in its left context), and (ii) - if applicable - in which relation it stands to this context (e.g. anaphor, instance, etc.).

3.2.3.2 Evaluation

High annotation quality and consistency is crucial for machine learning experiments. Evaluations and documentations of the annotation process help to estimate what performance can be expected from classifiers based on the respective data.

A common way of evaluating annotation guidelines is that of letting two or more annotators annotate the same texts independently and assess the extent to which they agree. Various measures have been used for this purpose, from simple percent agreement (the proportion of labels agreed on of the total number of instances to be labeled) to precision, recall and f-measure¹⁶ to more sophisticated measures like Cohen’s (1960) κ ¹⁷ (which accounts for agreement occurring randomly) and adaptations. For an overview of agreement measures, see e.g. Artstein and Poesio (2005).

For MUC, Hirschman et al. report an inter-annotator agreement of “84% precision and recall” (Hirschman et al., 1998, p. 4), which corresponds to 84% f-measure. Inter-annotator agreement rose to 91% f-measure (on a small number of test documents) after the task had been broken down into a step of identifying all markables in a text and then linking coreferring elements in a separate step.

¹⁴Or each referent mentioned more than once, respectively.

¹⁵In the case of cataphors or bound pronouns, this ‘antecedent’ could be found in the right context instead of the left context. In the case of multiple antecedents (e.g. *multiple phrases* in ARRAU, *split_antecedent* in TüBa-D/Z), the expression is annotated with its antecedents’ IDs.

¹⁶The same measures are used for evaluating classifiers. Definitions are given in Section 4.3.1, definitions 108, 109 and 110.

¹⁷ κ is calculated as follows: $\kappa = \frac{p_o - p_c}{1 - p_c}$, where p_o is the proportion of observed agreement and p_c the proportion of agreement by chance (i.e. as if the events were independent). The resulting value lies between 0 and 1. The closer κ is to 1, the more consistent the annotations are. According to Carletta (1996), $\kappa > .8$ is considered “good reliability”; $0.67 < \kappa < .8$ “allowing [for] tentative conclusions” (Carletta, 1996, p. 252).

Nissim (2006) reports κ -values of .902 for distinguishing *old* vs. *med/new* (based on 1,502 instances marked by two annotators).¹⁸

Hovy et al. report “average agreement scores between each annotator and the adjudicated results [of] 91.8%” (Hovy et al., 2006, p. 59) for OntoNotes. Markert et al. (2012) report on the annotation of information status of 26 texts from the *Wall Street Journal* portion of OntoNotes. The annotation makes use of OntoNotes’ original coreference annotation and is carried out by 3 annotators. Out of 5,905 NPs from the syntactic annotation, 1,499 were pre-marked as *old* using OntoNotes’ coreference annotation, “leaving 4406 potential mentions for annotation and agreement measurement” (Markert et al., 2012, p. 797). A pairwise evaluation of annotators’ agreement yields percentage values between 86.3% to 87.5% for a coarse-grained 4 category distinctions (values for Cohen’s κ between 0.747 and 0.773) and between 85.3% and 86.6% for a finer-grained 9 category distinction (κ between 0.773 and 0.801). For the category *old* only, they report κ values between 0.793 and 0.832.

For ARRAU, Poesio and Artstein (2005) and Poesio and Artstein (2008) report on evaluation experiments set up as follows: “multiple annotators (as many as 20) worked independently on the same text, and formal reliability measures such as a (Krippendorff, 1980) were used to compare the annotations and identify easy and difficult parts of the task; agreement on anaphoric chains was in the range of $\alpha \approx 0.6-0.7$ ” (Poesio and Artstein, 2008, p. 1171).

For TüBa-D/Z, Versley (2006) reports an f-measure of 83% (85% after mapping spans to nodes) for the full coreference task. In a detailed analysis, separating the referring expressions by their semantic class, he finds agreement is higher on NPs referring to persons and organizations than on NPs referring to temporal entities, events, objects, and locations.

For PCC, Krasavina and Chiarcos (2007) report κ values of 0.61-0.77 for 19 texts annotated by 2 annotators using the core scheme of PoCoS.¹⁹ As to information status, Ritz et al. (2008) report κ values of 0.55 for the extended scheme (9 different labels plus non-referring) and 0.60 for the core scheme (3 different labels plus non-referring) between 2 annotators (220 NPs). The commentary texts were chosen for the purpose of discourse structure research. They express subjective views on events and topics introduced elsewhere in the newspaper and frequently refer to entities known to locals at the time of publication (politicians, buildings, etc.). A reconstruction thus requires detailed background knowledge.

For DIRNDL, Riester et al. (2010) report κ values of 0.78 for a six-category core scheme, and 0.66 for the entire scheme. These numbers are based on 1,149 DPs/PPs²⁰ labeled by two annotators.

Obviously, these evaluations are not comparable: they use different measures, different

¹⁸For the distinction of old vs. mediated vs. new, results are reported as .845, and .788 for an even more fine-grained distinction. Note that instances tagged *non-applicable* (idioms, expletives, parsing errors etc.) and *not understood* (where the annotator did not fully understand the text) were excluded beforehand (Nissim, 2006).

¹⁹Their experiments with the English adaption of the scheme yielded κ values of 0.71-0.96 on 8 texts of the RST Discourse Treebank (*Wall Street Journal* articles). 6 annotators had annotated these texts with pair-wise overlapping portions.

²⁰Riester uses the term determiner phrase (DP) for referring expressions; PP’s are included due to fusion of preposition and definite determiner in German.

numbers of annotators, and different numbers of instances. Some of the evaluations are poorly reported, e.g. information on the number of annotators or the number of instances annotated are missing. What *can* be read from these studies, however, is that there is a considerable portion of cases that human annotators disagree on.

A discussion of annotation quality should also take into account other factors, like reusability and comparability. The reuse of an existing corpus is facilitated if it is provided with additional layers of annotation (e.g. syntactic structures, genericity information, semantic properties, named entity types etc.). This ensures that the basic information (e.g. NP boundaries in the case of this work) is the same for all users, which makes comparisons between studies possible. Beyond that, comparability between schemes is facilitated by fine-grained categories that (at least in parts) correspond to each other: for instance, assume that scheme 1 distinguishes categories A_1 and B_1 . Further assume that, for reasons of theoretical modelling, someone suggested that a subset S of A_1 be grouped with B_1 instead of A_1 . Then, it would be advisable if the revised version scheme 2 defined *three* categories D_2 , E_2 and F_2 with $D_2=S$, $E_2=A_1 \setminus S$ and $F_2=B_1$, so that experiments according to the new theory can be carried out (with E_2 replacing the old category A_1 and $D_2 \cup F_2$ replacing the old category B_1), while at the same time, a categorization according to the old scheme ($D_2 \cup E_2$ for A_1 and F_2 for B_1) is available.

On a meta level, efforts towards better comparability will hopefully lead to a distinction of essential vs. less essential, or uncontroversial vs. controversial categories.

3.2.4 A Critical Assessment of Existing Corpora

This section provides a summary of the strengths and limitations of existing resources. Criteria include corpus size, theory-foundedness of the notion of coreference/discourse-givenness, technical modeling, and consistency of annotation. As a general picture, fundamental improvements have been made.

Corpus Size

An overview of corpus sizes is given in Table 3.1 in Section 3.1. They range roughly from several thousand NPs (MUC, PCC and DIRNDL) to nearly 374,000 NPs (TüBa-D/Z). In general, classifiers benefit from larger amounts of data, such as provided in OntoNotes or TüBa-D/Z. After all, the corpus has to be divided into subsets for training, development and testing, respectively. Ng (2011) suggests a 60-20-20 splitting, i.e. 60% of the data is used for training purposes, and 20% each for validation and the final testing (or 80-10-10, alternatively). The effect of the amount of data used for training will be shown in Section 5.3.1.

Definition of Coreference

Cases uncontroversial across different corpora are cases of coreference between specific singulars and between specific definite plurals. However, there is a large number of sub-phenomena in coreference and phenomena close to coreference (see Sections 3.2.1 and 3.2.2 for examples and proportions). In the optimal case, the annotation distinguishes these (sub-)phenomena, so that a user of the corpus can selectively access instances of each of them.

The definition of coreference in early resources deviated in parts from the strict definition presented above. Since then, the conception of markables and relations has been refined: issues in the MUC annotation scheme listed under ‘further directions’²¹ and those criticized e.g. by van Deemter and Kibble (2000) and Kibble and van Deemter (2000) have been solved in succeeding resources. As to markables, for instance, OntoNotes and ARRAU allow markables that are not NPs; ARRAU additionally allows discontinuous markables. As to relations, the IDENT relation was used in MUC for coreference as well as appositions; it also included function-value assertions and predicative constructions. In OntoNotes, separate relations (IDENT and APPOS) were used for coreference vs. apposition; function-value assertions and predicative constructions are not included. Finer-grained distinctions are made in ARRAU (labeling of generic expressions) and TüBa-D/Z (including the relations coreferential vs. anaphoric vs. bound). This shows that the coreference definition in corpora is converging to the theoretical definition presented above. One phenomenon that remains hard to annotate is the entity-attribute relation. Issues that have not been solved satisfactorily yet in existing resources include time-dependence and predication.

Advanced Technical Modelling

Technical solutions have been implemented that enable the annotation of discontinuous markables (e.g. in ARRAU and PCC), directed edges (e.g. anaphors vs. cataphors in TüBa-D/Z and PCC), and ambiguities (e.g. in ARRAU and DIRNDL²²). Increasingly, synergy effects of multi-layer corpora are being used: constituents annotated at the syntactic layer form markables for information status annotation, e.g. in PCC. No syntactic annotation existed for the MUC corpora. The annotation rule for constituents was lax: spans did not have to include determiners, for instance. The MUC evaluation scripts equally accepted NPs with or without determiners to accommodate the lax definition. The downside of a lax definition is less consistent annotation. In OntoNotes, there are rarely differences between NPs on the syntactic level and markables on the coreference level.²³ However, as a consequence of reusing syntactic NPs as markables, in TüBa-D/Z, definite NPs with a fusion of the determiner and preposition do not include their definiteness feature (see Figure 3.3, Section 3.2.1 for details of the analysis), as it is attached to the preposition outside the NP.

Quality Assurance

Consistency of annotation is important for any corpus research, in particular for applications in machine learning. It is increasingly provided for: annotation schemes are tested

²¹Hirschman and Chinchor’s (1997) suggestions for further directions are:

- “1) coreference to cover clause (verbal) level relations
- 2) a method for handling discontinuous elements, including conjoined elements
- 3) a distinction between function/type coreference and instance coreference, which has caused some problems with the unintended merging of coreference chains
- 4) set/subset coreference, part/whole and other kinds of coreference” (p. 3).

²²For technical reasons, one label is created per antecedent link.

²³Verbal antecedents of event anaphors form an exception, of course.

for agreement among annotators. Studies on inter-annotator agreement are reported in more and more detail (see Section 3.2.3.2; ARRAU is a positive example here). Unfortunately, evaluation standards have evolved only recently, so earlier work cannot be easily compared.

As a sidenote, figures of inter-annotator agreement should never be emphasized over theory-foundedness: the downside of inter-annotator-agreement orientation is that more superficial definitions may seem superior judging just by the numbers. For instance, the notion of specificity in OntoNotes does not coincide with that in the literature, but can obviously annotated more consistently and with higher agreement.

To sum up, each of the resources has implemented improvements over earlier resources. This has led to corpora with annotations that are coming closer to the theoretical notion of coreference presented above.

Chapter 4

Related Work and Technical Background

Several different sets of features and classification algorithms have been employed in previous approaches. This chapter will give an outline of the data and categories used, the variety of features and algorithms that have been proposed, the classification experiments that have been carried out, and the results reported for these experiments. In addition, this chapter contains technical background information on the procedure of experiment evaluation (Section 4.3.1).

4.1 Data and Categories

An overview of the corpus data and categories used in the different approaches is given in Table 4.1. Details on these corpora and categorizations have been described in Chapter 3. Features, methods and results will be presented in more detail in the following sections. The overview shows that large parts of the work in the field of discourse-givenness classification has been carried out on MUC corpora or their successor ACE. A shift towards the classification of information status, rather than just discourse-givenness, can be observed in recent years.

author	category	corpus	agreement	method*
(Ng and Cardie, 2002)	<i>anaphoric vs. non-anaphoric</i>	MUC-6 and MUC-7	84% f-measure	J48, Ripper
(Uryupina, 2003)	\pm <i>discourse_new</i>	MUC-7	84% f-measure	Ripper
(Hempelmann et al., 2005)	<i>given, inferrable, new</i>	textbook texts	$\kappa=.72$	logistic regression
(Nissim, 2006)	<i>old, mediated, new</i>	Switchboard	$\kappa=.845$	J48
(Denis and Baldrige, 2007)	<i>anaphoric vs. non-anaphoric</i>	ACE	n.a.	ILP
(Ng, 2009)	<i>anaphoric vs. non-anaphoric</i>	ACE	n.a.	MaxEnt
(Uryupina, 2009)	\pm <i>discourse_new</i>	MUC-7	84% f-measure	SVM, Ripper
(Rahman and Ng, 2011)	<i>old, mediated, new</i>	Switchboard	$\kappa=.845$	SVM
(Zhou and Kong, 2011)	<i>anaphoric vs. non-anaphoric</i>	ACE	n.a.	label propagation with polynomial kernel
(Cahill and Riester, 2012)	<i>old, mediated, new, other</i> ¹	DIRNDL	κ^2	CRF
(Markert et al., 2012)	<i>non-mention, old, mediated, new</i> ³	OntoNotes	$\kappa=.747$ to $.773$	J48, SVM, ICA

Table 4.1: Overview: Approaches to Identifying Information Structure

- ¹. The more fine-grained versions of the label set are: *given, situative, bridging, unused, new, generic, expletive; given-pronoun, given-reflexive, given-noun, situative, bridging, unused-known, unused-unknown, new, generic, expletive; given-pronoun, given-reflexive, given-epithet, given-repeated, given-short, situative, bridging, unused-known, unused-unknown, new, generic, expletive*.
- ². The evaluation was carried out on an earlier version of the scheme, which distinguishes 21 categories, resulting in $\kappa=.66$; $\kappa=.78$ for the 6 top-level categories *given, situative, bridging, accessible, indefinite, other* (Riester et al., 2010).
- ³. *Non-mention* stands for nonreferring expressions. Sub-categories of *mediated* include *knowledge, synt, aggregate, func, comp and bridging*. κ for the fine-grained category set ranges from $.773$ to $.801$.

* Algorithms are discussed in Section 4.3.

Abbreviations: J48 - WEKA's implementation of C4.5 decision trees (according to (Witten and Frank, 2005); see discussion in Section 4.3), Ripper - repeated incremental pruning to produce error reduction, ILP - integer linear programming, SVM - support vector machines, CRF - conditional random field, ICA - iterative collective classification.

4.2 Features

The features proposed in different approaches make use of various levels of linguistic processing, such as the (tokenized) surface form, morphologic and syntactic analysis, comparison to the previous context, salience ranking, classification of named entity types, and lexical or distributional information. Several approaches use features that are variants of previously suggested features, e.g. a feature combined of several other features or a different coding (e.g. several boolean features instead of one numeric or categorial feature). For this reason, the features will be introduced in groups of what they describe.¹

4.2.1 Properties of NPs

Regarded in isolation, an NP can be described in terms of its

1. surface form (the words it consists of),
2. length (the number of tokens or characters it contains),
3. spelling (capitalization of some or all characters, whether it contains digits or special characters),
4. morphological features (number, gender, person),
5. syntactic form and structure (pronominalization, existence and form of determiners, use of common or proper nouns, parts of speech contained, modification and complexity), as well as
6. its semantic class (e.g. NE type).

Some of the following approaches use features for coreference resolution which they do not use for the detection of anaphoric expressions. These will not be listed.

1. Surface Form

Nissim (2006)² reports on running experiments including the NP string, hoping it would help with the classification of “general mediated instances (common knowledge entities), such as ‘the sun’, ‘people’, ‘Mickey Mouse’, and so on” (Nissim, 2006, p. 97). However, she observes a negative effect on the classifier’s performance and excludes the feature from the model. Rahman and Ng (2011) and Markert et al. (2012) use all unigrams (i.e. words) appearing in any mention in the training set. Markert et al. (2012) additionally have a feature capturing whether the mention is modified by a comparative marker (‘another’, ‘such’, ‘similar’ etc.) from a list of 10 markers.

2. Length

Very long NPs tend to be first mentions. The length of an NP, measured in tokens, is used by Nissim (2006), Denis and Baldridge (2007), Uryupina (2009), Rahman and Ng (2011), Cahill and Riester (2012) and Markert et al. (2012). Cahill and Riester (2012) include a discretized version of the feature ‘length’: they have boolean features for phrases consisting of less than 2, less than 5 and less than 10 words.

¹Groups are sequentially numbered, numbering in Section 4.2.2 continues from Section 4.2.1.

²Nissim’s (2006) feature set has been reused and extended by Rahman and Ng (2011) and Markert et al. (2012).

3. Spelling

In English, capital letters are contained only in names and abbreviations, not in common nouns (in German, common nouns are also capitalized). Names are specific mentions, although they do not need a definite determiner. Many first mentions of abbreviations contain the spelt out version. Ng and Cardie (2002)³ have a boolean feature for NPs that are entirely in uppercase. Uryupina (2003; 2009) uses features capturing the proportion of upper or lowercase letters, of digits and of special characters.

4. Morphological Features

Ng and Cardie (2002), Rahman and Ng (2011), Cahill and Riester (2012) and Markert et al. (2012) use number as a feature ('singular'/'plural'/'unknown'). Zhou and Kong (2011) use boolean features stating whether the NP is in the singular, and whether it is a male/female pronoun. Information on number, person and gender is used by Uryupina (2009).

5. Syntactic Form and Structure

An NP's pronominalization, the existence and form of determiners, the use of common or proper nouns, and information on parts of speech⁴, modification and complexity have been used in various codings and combinations.

Ng and Cardie (2002), Ng (2004), Hempelmann et al. (2005), Denis and Baldrige (2007), Zhou and Kong (2011) and Cahill and Riester (2012) use a boolean feature representing whether or not the NP in question consists of a pronoun. Cahill and Riester (2012) additionally use the pronoun type (e.g. 'demonstrative'). Information on pronominalization is contained in an NP's parts of speech. Uryupina indirectly accesses this information using the parts of speech of the NP's head word (Uryupina, 2003), or the parts of speech of all words in the NP (Uryupina, 2009), respectively. Nissim (2006), Rahman and Ng (2011) and Markert et al. (2012) have a feature 'NP type' which combines several pieces of information and can take any of the values 'pronoun', 'common', 'proper', or 'other'. Regarding determiners, Ng and Cardie (2002), Ng (2004), Hempelmann et al. (2005) and Zhou and Kong (2011) use a boolean feature coding whether or not an NP has a definite determiner. Ng and Cardie (2002) and Zhou and Kong (2011) add a feature for demonstrative NPs. Uryupina (2003; 2009) as well as Nissim (2006), Rahman and Ng (2011) and Markert et al. (2012) have a compound feature specifying the determiner type (in the latter three works, the feature's value set contains 'definite', 'demonstrative', 'indefinite', 'bare', 'possessive', and 'not applicable'). Ng and Cardie (2002) have boolean features for indefinite NPs (starting with the determiner 'a' or 'an'), and quantified NPs (starting with 'every', 'some', 'all', 'most', 'many', 'much', 'few' or 'none'), as well as a combined categorial feature 'article' (definite/indefinite/quantified). Other features cover whether the NP is a proper noun, possessive, bare singular, bare plural, has a pronominal modifier, premodifier, postmodifier, contains special nouns (comparative noun or premodified by a superlative). They also use patterns of parts of speech (e.g. 'the PN N', etc.). Cahill and Riester (2012) use the determiner type (e.g. 'definite'), and the type of the head noun (e.g. 'common'), where applicable. Boolean features are used to represent whether the phrase contains a compound noun and whether it contains a time expression. The parts of speech of the leftmost and rightmost token, respectively, are

³Ng and Cardie's (2002) feature set has been reused in Ng (2009).

⁴Pronominalisation and the presence or absence of determiners is of course evident from the list of an NP's parts of speech.

also recorded. Features describing an NP’s modifiers include Uryupina’s “heuristics for restrictive postmodification” (Uryupina, 2003, p. 83) and her (2009) (not further specified) features covering “pre- and postmodification”. Markert et al. (2012) use a feature representing the presence of adjectives, one representing the presence of adverbs or adverbs combined with a comparative (see number (1.) above). Zhou and Kong (2011) use a feature capturing whether the NP embeds another NP. Ng and Cardie (2002) and Cahill and Riestler (2012) use a boolean feature that is true if the NP is a conjunction. As a measure of the NP’s complexity, Cahill and Riestler (2012) record whether or not the respective phrase contains more than 1 DP⁵ and 1 NP and count the number of DPs and NPs contained in the respective NP, the number of top category children, as well as the depth of the syntactic phrase (with and without unary branching, respectively). The syntactic shape is described in another feature, e.g. ‘apposition with a determiner and attributive modifier’. Also included are features counting cardinal numbers and year phrases. A boolean feature is used to flag phrases that do not have complete parses.

6. Semantic Class

Ng (2004) and Uryupina (2009) make use of the semantic class of the NP, i.e. the named entity type (e.g. person, organization, location, date, time, etc.). Markert et al. (2012) also use the semantic class (one of 12 classes like location, organization, person, date, money, percent etc.), which is derived from the OntoNotes entity type annotation and an “automatic assignment of semantic class via WordNet hypernyms for common nouns” (Markert et al., 2012, p. 800). Ng and Cardie (2002) use titles and positions. Cahill and Riestler (2012) use the adverbial type (e.g. locative), and also the number of titles (“# Labels/titles” as one of the countable features, p. 234).

4.2.2 NPs’ Relations to their Context

An NP’s relation to its context⁶ (sentence, discourse, and corpus) can be characterized by

7. its grammatical function (or, as a heuristic, position within the sentence),
8. words (or parts of speech of the words) occurring in its local context (i.e. the sentence or a window of fixed size),
9. its salience,
10. its occurrence in certain constructions (predications, appositions, modal constructions, or under negation),
11. agreement, identity with (or similarity to) other mentions in the discourse, as well as the distance to identical/similar mentions,
12. the probability for its occurring as a definite description.

7. Grammatical Function and Position

The feature ‘grammatical function’ is used by Nissim (2006), Uryupina (2009), Rahman

⁵Riestler uses the term Determiner Phrase (DP) for referring expressions.

⁶Item numbers follow up the numbering in Section 4.2.1.

and Ng (2011), Cahill and Riester (2012) and Markert et al. (2012); Nissim specifies the value set as ‘subject’, ‘passive subject’, ‘object’, ‘pp’, ‘other’. Ng and Cardie (2002) use the NP’s position (‘first sentence’, ‘first paragraph’, ‘header’). Zhou and Kong (2011) use the NP’s semantic role: whether the NP has a semantic role, whether it is ‘Arg0’ (agent), whether it is ‘Arg0’ (agent) of the main predicate of the sentence. Additionally, they use its position in the sentence: they mark NPs that are the first NP in the sentence.

8. Local Context

Zhou and Kong (2011) take into account whether the NP is embedded in another NP. They also count ‘Forward and Backward Distance’, i.e. “the distance between the current NP and the nearest [forward/] backward clause, indicated by coordinating words (e.g. that, which)” (p. 38). Rahman and Ng (2011) and Markert et al. (2012) use partial parse trees (the respective node’s parent and sibling nodes without the lexical leaves). Cahill and Riester (2012) use the label of the highest syntactic node that dominates the phrase.

9. Saliency

Uryupina (2009) tests various saliency rankings (without giving further details), later observing that they “show virtually no performance gain over the baseline” (Uryupina, 2009, p. 117).

10. Sentence Construction

Ng and Cardie (2002), Uryupina (2003), Ng (2004) and Zhou and Kong (2011) use a feature stating whether or not the NP is used as an apposition. Ng and Cardie (2002) also have a boolean feature for predicate nominal constructions. Uryupina, in her later work (2009), adds features for “copula, negation, [and] modal constructions” (p. 117). Cahill and Riester (2012) use one feature indicating predications and one indicating appositions.

11. Agreement, Identity and Similarity

Number and gender agreement is used by Ng (2004). Ng and Cardie (2002) use the boolean features ‘string match’ (after discarding determiners) and ‘head match’, ‘alias’ and ‘subclass’. Uryupina (2003) computes the distance to the previous NP with the same head (if existent), measured in NPs and in sentences. In her (2009) classification, she additionally uses the (self-explanatory) boolean feature ‘same head exists’. Ng (2004) has four identity features. There are three boolean features, one for pronouns, one for proper names, and one for non-pronominal NPs each, capturing whether (after the deletion of determiners and demonstrative pronouns) there is a matching NP in the left context of the NP in question. Additionally, there is a numeric value measuring the distance between markables in sentences. Nissim (2006) has three features of group (11): a numeric feature ‘full previous mention’ counting the number of identical NPs in the left context; its categorial version ‘mention time’ (with the values ‘first’, ‘second’, or ‘more’), and ‘partial previous mention’ (with the values ‘yes’, ‘no’, ‘not applicable’). The latter is explained only by example (“for example, ‘your children’ would be considered a partial previous mention of ‘my children’ or ‘your four children’. The value [...] ‘non-applicable’ [...] is mainly used for pronouns” (Nissim, 2006, p. 97)). Denis and Baldrige (2007) use a feature capturing whether there is a previous mention with a matching string, as well as the number of preceding mentions. Rahman and Ng (2011) and Markert et al. (2012) reuse Nissim’s (2006) features; Markert et al. (2012), however, change the numeric feature ‘full previous mention’ into a categorial one (‘yes’, ‘no’, ‘NA’) and add features

that count the number of times an NP has been partially mentioned previously, as well as the number of times an NP’s content words have been mentioned previously. Zhou and Kong (2011) have one boolean feature indicating whether there exists an NP consisting of the same string.

Cahill and Riestler (2012) have a boolean feature stating whether the head noun appears (partly or completely) in the previous 10 sentences. Additionally, they use GermaNet⁷ (Hamp and Feldweg, 1997) to calculate (i) the distance of the head noun’s synset to the root node (assuming that a term that is more general is more likely to be used generically), and (ii) the sum and the maximum of semantic relatedness (Lin, 1998)⁸ of the head noun to its immediately preceding neighbour phrases (using GermaNet Pathfinder (Finthammer and Cramer, 2008) for calculations). Zhou and Kong (2011) have a boolean feature ‘NameAlias’ capturing whether the NP and any of the NPs in the context are name aliases or abbreviations of the other. They also use a feature stating whether there is an NP in the context which agrees with the current NP in word sense (annotations of WordNet senses are used for this).

Hempelmann et al. (2005) use latent semantic analysis (LSA) and ‘span’, a variant of LSA, to measure the similarity between a text item (probably an NP) and its context. LSA is based on vectors representing word cooccurrences, drawn from a large text corpus. The similarity is computed as the cosine of the angle between two vectors, each representing a piece of text (the NP or the context, respectively).

Definiteness Probability

Uryupina (2003) includes “definite probability features”, four features measuring the proportion of an NP’s occurrences as a definite description (compared to its occurrences as an indefinite description, i.e. as a bare noun or with the determiner *a* or *an*). These features are based on a web count and calculated as shown in (99).

$$(99) \quad p_1 = \frac{\#''the Y''}{\#Y}; p_2 = \frac{\#''the H''}{\#H}; p_3 = \frac{\#''the Y''}{\#''a Y''}; p_4 = \frac{\#''the H''}{\#''a H''},$$

where Y stands for the full NP (without the determiner), H for its head noun.

4.3 Algorithms

Various classification algorithms have been employed. They are briefly characterized in the following. More details can be found, for instance, in Witten and Frank (2005), Kotsiantis (2007), and Ng (2011) among others.

Basically, the learning methods can be distinguished into supervised and semi-supervised methods. Within the group of supervised methods, symbolic methods (decision tree learners and rule learners), functions and sequential learners can be distinguished.

Symbolic methods produce models that represent the conditions for an instance to belong to a certain class as a logic combination of the input features. The resulting models are commonly considered intuitively interpretable by humans, like decision trees or sets of rules.

⁷GermaNet, the German version of WordNet, is a lexical resource in the form of a net. Nodes represent word senses, edges represent relations, like antonymy, hyperonymy/hyponymy etc.

⁸Lin distinguishes semantic relatedness from semantic similarity. The former is calculated based on paths in an ontological resource, the latter is calculated based on word distributions in texts.

Decision trees are hierarchical models, where leaves represent class values, inner nodes represent features, and edges represent values of these features.

Figure 4.3 illustrates what a decision tree for discourse-givenness could look like.

```

head_previously_mentioned>0
| determiner=definite:given
| determiner!=definite:new
head_previously_mentioned<=0
| form=pronoun:given
| form!=pronoun
| | determiner=none
| | | position=initial:given
| | | position!=initial:new
| | determiner=indef:new
...

```

Figure 4.1: Example Decision Tree for Discourse-Givenness.

The tree is read as follows (‘|’ stands for a branching): If the head of the respective NP has been previously mentioned before (more than 0 times), and if it has a definite determiner, it is classified as *given*, otherwise it is classified as *new*. If the head has not been mentioned, and if it is a pronoun, it is classified as *given*; NPs other than pronouns, if they do not have a determiner and occur in initial position, are classified as *given*, in non-initial position as *new*; if they have an indefinite determiner, they are classified as *new*, etc.

The method of constructing decision trees is described by the following piece of pseudo-code taken from Kotsiantis (2007), p. 252⁹:

```

"Check for base cases [technical requirement due to recursion, N.B.]
  For each attribute a
    Find the feature that best divides the training data
    [using a measure] such as information gain [the reference
    provided is Hunt et al. (1966)] from splitting on a
  Let a_best be the attribute with the highest normalized information gain
  Create a decision node node that splits on a_best
  Recurse on the sub-lists obtained by splitting on a_best and add those
  nodes as children of node"

```

Figure 4.2: Pseudo-code for Decision Tree Construction (Kotsiantis, 2007, p. 252).

Algorithms for learning decision trees include C4.5 (Quinlan, 1993). According to Witten and Frank (2005), *J48* in their machine learning toolkit WEKA¹⁰ represents a reimplementation of C4.5. Moore et al. (2009), however, in a study on the performance of the decision tree algorithms J48, C4.5 and C5.0 (“an updated commercial version of C4.5”,

⁹Kotsiantis uses the term *attribute* in the sense that *feature* has been used here, and *feature* for feature values (or ranges thereof).

¹⁰Waikato Environment for Knowledge Analysis (Witten and Frank, 2005), <http://www.cs.waikato.ac.nz/~ml/weka/index.html>

p. 185) on a range of different data sets, come to the conclusion that J48 “performs much more similarly to C5.0 than to C4.5” (p. 186). Ng and Cardie (2002) report they used C4.5 in their experiments; Nissim (2006) used J48. Markert et al. (2012) also use J48 in comparison to ICA; the results for the class *old* are only slightly better using ICA, they are substantially better for the classes *mediated* and *new*.

A decision tree can be translated into a set of rules (Quinlan, 1993; Kotsiantis, 2007). The set of rules corresponding to the excerpt of the decision tree in Figure 4.3 is shown in Figure 4.3.

```
(head_previously_mentioned>0) and (determiner=definite) => class=given
(head_previously_mentioned>0) and (determiner!=definite) => class=new
(head_previously_mentioned<=0) and (form=pronoun) => class=given
(head_previously_mentioned<=0) and (form!=pronoun) and (determiner=none) and
(position=initial) => class=given
(head_previously_mentioned<=0) and (form!=pronoun) and (determiner=none) and
(position!=initial) => class=new
(head_previously_mentioned<=0) and (form!=pronoun) and (determiner=indef) =>
class=new
...
```

Figure 4.3: Example Rules for Discourse-Givenness.

The method for constructing a set of rules from a data set is described by Kotsiantis (2007) with the following piece of pseudo-code:

```
"On presentation of training examples [...]:
1. Initialise rule set to a default (usually empty, or a rule assigning
   all objects to the most common class).
2. Initialise examples to either all available examples or all examples
   not correctly handled by rule set.
3. Repeat
   (a) Find best, the best rule with respect to examples.
   (b) If such a rule can be found
       i. Add best to rule set.
       ii. Set examples to all examples not handled correctly by rule
           set.
until no rule best can be found
(for instance, because no examples remain)."
```

Figure 4.4: Pseudo-code for Rule Learners (Kotsiantis, 2007, p. 253).

Algorithms for learning rules include Ripper (‘Repeated Incremental Pruning to Produce Error Reduction’ (Cohen, 1995); *JRip* in WEKA). Ripper has been applied by Ng and Cardie (2002) (they report C4.5 yielded better results, though). Uryupina also used Ripper arguing that “[f]irst, we need an algorithm that does not always require all the features to be specified. [...] and [s]econd, we want to control precision-recall tradeoff” (Uryupina, 2003, p. 83), both in her (2003) and (2009) work.

The second group of supervised methods is functions. These methods consist in learning weights for the input features during training. The set of weights is optimized regarding a certain criterion (the cost function) characteristic to each of the different approaches (ILP has the additional restriction that one or more of the variables need to be integer values, i.e. not continuous values).

Assume a set of m training instances in the form of a matrix, with vectors $x^{(1)}, \dots, x^{(m)}$. Each vector contains the values of the input features $x_1^{(i)}, \dots, x_n^{(i)}$ (n values in numeric or binary coding), and $y^{(i)}$ contains the class (0 or 1). In Logistic Regression, the hypothesis which class a given instance belongs to can be formulated as shown in (100) (formulae adapted from Ng (2011) for a uniform representation).¹¹

$$(100) \quad h_{\Theta}(x) = g(\Theta^T x), \text{ where } \Theta \text{ is the vector of weights to be learnt, and } g \text{ a transformation function with the additional requirement } 0 \leq g \leq 1$$

For classification purposes, where the value of y needs to be either 0 or 1, the following function is used:

$$(101) \quad g(y) = \frac{1}{1+e^{-y}}$$

Ng (2011) gives the cost function as in (102). Parameters Θ need to be fit to minimize this cost function.

$$(102) \quad J(\Theta) = -\frac{1}{m} \sum_{i=1}^m (y^{(i)} \log h_{\Theta}(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_{\Theta}(x^{(i)})))$$

Logistic Regression is applied to the discourse-givenness classification task by Hempelmann et al. (2005).

In comparison to Logistic Regression, the prediction condition for SVMs is given in (103) (Ng, 2011).

$$(103) \quad y = 1 \text{ is predicted if } \Theta^T x \geq 0$$

For SVMs (Vapnik, 1995) without a kernel (also termed ‘with a linear kernel’), Ng (2011) gives the optimization criterion in 104, with the additional parameter λ to control the bias/variance tradeoff¹².

$$(104) \quad \min_{\Theta} \frac{1}{\lambda} \sum_{i=1}^m (y^{(i)} \text{cost}_1(\Theta^T x^{(i)}) + (1 - y^{(i)}) \text{cost}_0(\Theta^T x^{(i)})) + \frac{1}{2} \sum_{j=1}^n \Theta_j^2$$

with $\text{cost}_1(\Theta^T x^{(i)}) = -\log(h_{\Theta}(x^{(i)}))$ and $\text{cost}_0(\Theta^T x^{(i)}) = -\log(1 - h_{\Theta}(x^{(i)}))$

Here, minimizing the cost corresponds to maximizing the margin between positive and negative instances. In Figure 4.5, for instance, model A would be preferred over model B.

SVMs can learn non-linear functions with the help of different kernels (an additional parameter again allows for a control of the bias/variance tradeoff), see (105) for an example from (Ng, 2011).

¹¹Notation: T transpose of vector. Θ^T and x are combined with the dot product operator.

¹²*High bias* refers to a model’s property of not being complex enough to fit the data. *High variance* refers to a model’s property of overfitting the training data and thus potentially poor performance on unseen data. Small values for λ produce models with lower bias and high variance; large values for λ produce models with higher bias and low variance (Ng, 2011).

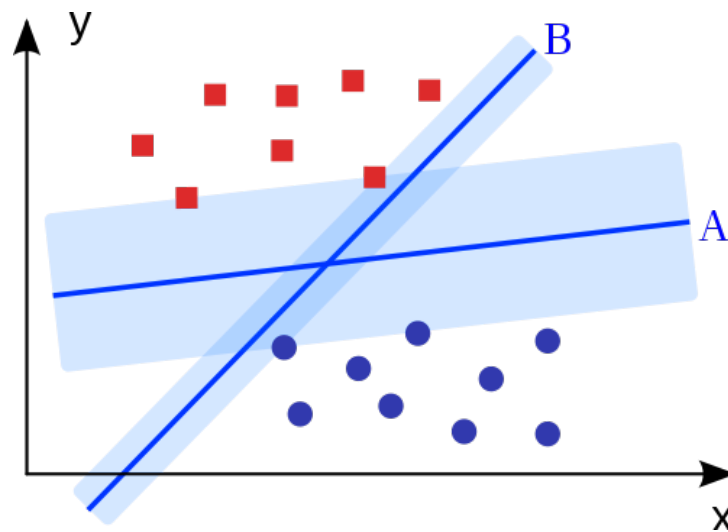


Figure 4.5: Support Vector Machines aim at maximizing the margin between instances of different classes (Source: http://de.wikipedia.org/wiki/Support_Vector_Machine, last access 6.02.2013)

(105) Predict ‘ $y = 1$ ’ if $\Theta_0 + \Theta_1x_1 + \Theta_2x_2 + \Theta_3x_1x_2 + \Theta_4x_1^2 + \Theta_5x_2^2 + \dots \geq 0$

See Witten and Frank (2005) or Ng (2011), for instance, for further details¹³, as well as for calculation methods solving the minimization problems formulated above.

Uryupina (2009) uses SVMlight, an SVM implementation in C (Joachims, 1999). Rahman and Ng (2011) use SVMs with a composite kernel (for handling the complex syntactic tree features). Markert et al. (2012) also use SVMlight (with a composite kernel); with better results for the class *old* in comparison to ICA (but lower results than ICA for the classes *mediated* and *new*).

MaxEnt (Maximum Entropy) models represent “multinomial logistic regression model[s.] [...] generaliz[ing] logistic regression by allowing more than two discrete outcomes”¹⁴, predicting the probability of each class value for a categorical class.

A maximum-entropy based classifier is used by Ng (2009).

In ILP (Integer Linear Programming; Nemhauser and Wolsey (1988)), the problem is formulated as an optimization task with the additional requirement that variables (or some of the variables) represent integers (i.e. are not discrete and not categorical).

Denis and Baldrige (2007) use ILP, claiming that it is “much more efficient than conditional random fields [a technique to be briefly discussed below, N.B.], especially when long-distance features are utilized” (p. 237).

Algorithms that sequentially label input data, i.e. take into account the features and the labels they predict for the elements’ context, include ICA (Iterative Collective Classifi-

¹³As for practical advice using Logistic Regression and SVMs, for instance, Ng suggests the use of Logistic Regression or SVMs without a kernel for learning tasks with $m \geq 50,000$ training instances and $1 \leq n \leq 1,000$ features as a rule of thumb.

¹⁴http://en.wikipedia.org/wiki/Multinomial_logistic_regression, last access 24.05.2013.

cation; Lu and Getoor (2003)) and CRF (Conditional Random Field; see Lafferty et al. (2001), and Klinger and Tomanek (2007) for a comparison to other statistical models). CRF is an undirected-graph based model and represents an extension of Logistic Regression to sets of interdependent variables.¹⁵ The most recent works employ methods for sequential tagging: Cahill and Rieger (2012) use CRF, and Markert et al. (2012) ICA. A semi-supervised method applied to discourse-givenness classification by Zhou and Kong (2011) is Label Propagation (Zhu and Ghahramani, 2002), where the “natural clustering structure in data is represented as a connected graph” (Zhou and Kong, 2011, p. 980). They describe this method as follows: each instance (labeled or unlabeled) is represented as a vertex. Edges between vertices are weighted by the similarity of the instances they represent. Similarity is modeled by a kernel; Zhou and Kong test a feature-based RBF (Radial Basis Function) kernel and convolution tree kernel. These weights are used to propagate labels from any vertex to neighboring vertices to “finally infer [...] the labels of unlabeled instances until a global stable stage is achieved” (Zhou and Kong, 2011, p. 980).

Kotsiantis’ (2007) comparison of different supervised classification techniques covers rule learners, decision tree learners, and SVMs. Table 4.2 represents an excerpt of his results, completed for logistic regression with the help of King et al. (1995).¹⁶

	rule learners	decision trees	SVM	logistic regression
dealing with discrete/binary/continuous attributes	*** (not directly continuous)	****	** (not discrete)	**** (p. 27)
accuracy in general	**	**	****	***
speed of learning with respect to number of attributes and number of instances	**	***	*	** (p. 27)
tolerance to highly interdependent attributes	**	**	***	*2
tolerance to irrelevant attributes	**	***	****1	****
model parameter handling	***	***	*	** (p. 22)
interpretability/comprehensibility of model	****	****	*	***

Table 4.2: Comparison of Learning Algorithms. ‘****’ represents best performance, ‘*’ worst performance (excerpt from Kotsiantis (2007), p. 263, ordered by relevance to this work; rightmost column completed according to King et al. (1995)).

¹ In contrast to this, Witten and Frank (2005) warn that SVMs with RBF kernels cannot deal effectively with irrelevant attributes (p. 234).

² CRF is an extension to cover for sets with interdependent attributes, see above.

¹⁵http://en.wikipedia.org/wiki/Logistic_regression, last access 24.05.2013.

¹⁶To the best of my knowledge, there is no comparable literature available for the other approaches.

4.3.1 Evaluation Measures

When having to choose the best performing classification model among a set of possible models, or when having to decide whether changes really bring an improvement in quality, one needs evaluation measures. This section is a compilation of the common practices as described, for instance, in Manning and Schütze (1999) (p. 268f.) and Ng (2011), relevant to the evaluation.

Naming conventions used for this purpose are shown Table 4.3: each instance has an actual class and a class as predicted by the classifier. Instances belonging to class 1 for which the classifier also predicts class 1 are called ‘true positives’, instances of class 0 for which the classifier predicts class 1 are called ‘false positives’ etc.¹⁷

		actual class	
		1	0
predicted class	1	TP	FP
	0	FN	TN

Table 4.3: Naming Conventions for Evaluation: True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN).

Some of the most commonly used measures for a classifier’s performance are error rate and accuracy: error rate is the proportion of wrong predictions among all predictions. Accuracy (also called *success rate*) is the proportion of correct predictions (see definitions (106) and (107)). These measures are two sides of the same coin, they sum to 100%, and the aim is, of course, to obtain a low error rate, i.e. high accuracy.

(106) DEFINITION *Error Rate*

$$error = \frac{\#false\ predictions}{\#predictions} = \frac{FP+FN}{TP+FP+FN+TN}$$

(107) DEFINITION *Accuracy* (also termed *Success Rate*)

$$acc = \frac{\#correct\ predictions}{\#predictions} = \frac{TP+TN}{TP+FP+FN+TN}$$

Accuracy works well for data sets with balanced class distributions, i.e. where each of the classes occur similarly frequently. However, there are data sets with a skewed class distribution, i.e. where one class occurs much more often than the other. In these cases, it can be misleading to rely on accuracy (or error rate, respectively) only. Ng (2011) illustrates this fact as follows with the example of a classifier predicting whether a patient’s tumor is malignant (i.e. cancerous) or benign, based on a range of examination values. Assume that 0.5% of all patients actually have cancer; an extremely skew distribution. A classifier always predicting that the patient does not have cancer would have an accuracy value of 99.5%. The high accuracy value suggests that the classifier is performing well, though it is obviously of no practical use, as none of the patients receives further treatment.

For that cause, another set of measures exists: precision and recall, along with the harmonic means of both measures, f measure. They are defined as shown in (108) to (110). Precision measures how many of the instances predicted positive actually are positive. Recall measures how many of the positive cases have been detected as being positive.

¹⁷Multi-way classification is discussed below.

The aim is to obtain high values for both Precision and Recall. The classifier always predicting ‘no cancer’ described above has a Recall value of 0%, which reflects its deficiency in detecting actually positive cases.

(108) DEFINITION *Precision*

$$P = \frac{TP}{\#predicted\ positive} = \frac{TP}{TP+FP}$$

(109) DEFINITION *Recall*

$$R = \frac{TP}{\#actual\ positive} = \frac{TP}{TP+FN}$$

(110) DEFINITION *F measure* (also termed *F score*)

$$F = \frac{2*P*R}{P+R} = \frac{2*TP}{2*TP+FP+FN}$$

Besides calculating precision, recall and f measure per class, there is also the option of weighted average of precision, recall or f measure: the weighted average is the sum of all per class values (either precision, recall or f measure), where each of the values is “weighted according to the number of instances with that particular class label”¹⁸, see definition in (111) (F is substitutable by P or R , respectively).

(111) DEFINITION *Weighted Average*

$$W_F = \sum_{i=1}^c \frac{N_i}{N} F_i$$
, with c the number of classes, F_i the f measure of class i , N the number of instances, and N_i the number of instances in class i .

In the work at hand, the class values are not equally distributed (for details, see Sections 5.3.1.1 to 5.3.1.4), but also not as skewed as in the tumor classification example above. For this reason, accuracy will be used as the measure for optimization, while precision, recall and f measure of the smaller class, as well as the weighted averages will be monitored.¹⁹

If the classification task is not binary, i.e. includes more than two classes, the measures are calculated as follows: for accuracy, a confusion matrix of all classes needs to be created. Accuracy is then calculated as the sum along the diagonal (correct predictions) divided by the total number of instances. Precision, Recall and F-measure are calculated per class: assuming we have three classes, *given*, *mediated* and *new*, and are calculating Precision etc. for the class *given*, then we consider *given* as class 1 and the other classes *mediated* and *new* as the complementary class 0 (i.e. *not given*) and proceed with the calculation as described above.

An interpretation of data sets involves comparisons. To this end, usually, one starts with assuming that the distribution of values of a certain feature follows the same pattern across different subsets of the data. This assumption is the so-called *null hypothesis*, which can then be put to the test by applying one of the following test statistics: for independent samples, Pearson’s chi-square (χ^2 , Pearson (1900), p. 165) with Yates’ correction (Yates, 1934) if needed; for dependent samples, McNemar’s (1947) test, also with Yates’ correction.

¹⁸<https://list.scms.waikato.ac.nz/pipermail/wekalist/2009-December/046789.html>, see also <http://weka.sourceforge.net/doc.dev/weka/classifiers/Evaluation.html>, last access 25.05.2013.

¹⁹Experiments with this work’s data have shown that optimization for f measure of the smaller class does not lead to substantially different results. I assume a distribution to be skewed if there is one class with a proportion in the lower single-digit percentage area or less.

The test statistics are defined as follows:²⁰

(112) DEFINITION Pearson’s χ^2

$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$, where n the number of cells in the contingency table, O_i is the observed frequency, and E_i the expected frequency (according to the null hypothesis). Yates’ correction: if any of $E_i < 5$, $O_i - E_i - 0.5$ is used instead of $O_i - E_i$.

(113) DEFINITION McNemar’s χ^2 with Yates’ correction

$$\chi^2 = \frac{(|FP - FN| - 0.5)^2}{FP + FN}$$

From χ^2 and the degrees of freedom, a p-value can be calculated (this is realized as a lookup in a pre-calculated table). This p-value represents the level of significance. The levels of significance are commonly defined as: highly significant ($p < 0.001$), very significant ($p < 0.01$), significant ($p < 0.05$).

Class distributions are compared using Pearson’s χ^2 (different sets are mutually exclusive). Classifier performances are compared using McNemar’s test (the same instance is classified by different classifiers). For this purpose, a contingency table is computed which contains for each classifier the correctly vs. incorrectly classified instances (see Table 4.4 for an example contingency table).

		classifier 1	
		corr	incorr
classifier 2	corr	1,600	150
	incorr	200	250

Table 4.4: Example Contingency Table (corr = correctly classified, incorr = incorrectly classified)

4.4 Experiments and Results

In this section, the results of previous work are presented. For each approach, a description of the category (terminology, values, and class distribution), as well as the data (and how it was processed, e.g. parsed, split into training and test set), features (by groups introduced in Section 4.2), experiment setup and results is given. The class distribution and results are shown in Table 4.4. Differences in class distributions can originate from differences between text types and from differences in annotation schemes: Hempelmann et al.’s (2005) textbook texts, for instance, contain a proportion of *given* expressions which is higher than in other texts, as the textbook texts are highly cohesive. Additionally, situationally evoked expressions are considered *given*. Similar to that, Nissim’s (2003) scheme defines all pronouns referring to the dialogue participants as *given*. Machine Learning methods generally tend to perform better on balanced data sets, i.e. data sets with classes that are nearly equally distributed.

²⁰Formulae adapted from http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test and http://en.wikipedia.org/wiki/McNemar%27s_test, respectively (last access 28.03.2013).

authors/experiments	class label (distribution)	P	R	F	acc	#NPs in set
(Ng and Cardie, 2002)	anaphoric (36.2%*)				86.1%	
MUC-6	not anaphoric (63.8%)					
MUC-7	anaphoric (26.8%*)				84.0%	Training: Dryrun Test: Formaleval
(Uryupina, 2003)	not anaphoric (73.2%)					(Dryrun) 3,710 (hold-out ¹)
MUC-7	-discourse_new (29.2%)	ø65.8%*	ø73.4%*	ø69.4%*	ø81.1%*	
(Hempelmann et al., 2005)	+discourse_new (70.8%)	ø88.5%	ø84.3%	ø86.3%	80%/74%/66% ²	478 (no training)
textbook	given (66.3%)					
	inferrable (24.3%)					
	new (9.4%)					
(Nissim, 2006)	old (48.3%*)	86.3%	81.5%	83.8%	93.1% ³	Train 40,865
Switchboard	med/new (51.7%*)	83.8%	88.1%	85.9%		Dev 10,565
old vs. med/new						Test 12,624
old vs. med vs. new						
	old (48.3%*)	94.1%	91.5%	92.8%	79.5% ³	
	med (36.8%*)	68.1%	87.6%	76.6%		
	new (14.8%*)	56.3%	22.3%	32.0%		
(Denis and Baldridge, 2007)	anaphoric (n.a.)				ø80.2%	train, test (original split)
ACE	-anaphoric (n.a.)					
(Ng, 2009)	anaphoric (n.a.)				n.a.	train, test (original split)
ACE	not anaphoric (n.a.)					
(Uryupina, 2009)	-discourse_new (34.1%*)	71.6%*	88.7%*	79.2%*	84.4%*	Train (Dryrun) 5,028
MUC-7	+discourse_new (65.9%*)	93.5%	82.3%	87.6%		Dev (Train) 976 Test (Formaleval) 3,375
(Rahman and Ng, 2011)	old (51.7%) ⁴	95.2%	93.0%	94.1%	82.2%	Train 60,703
Switchboard	med (34.2%)	70.9%	89.1%	79.0%		Test 8,982
(Zhou and Kong, 2011)	new (14.1%)	71.5%	34.4%	46.5%		
ACE	anaphoric (n.a.)				71.8-76.2%	train, test (original split)
	non-anaphoric (n.a.)				81.3-84.4%	

Table 4.5: State of the Art Classification Results (part I)

authors/experiments	class label (distribution)	P	R	F	acc	#NPs in set
(Cahill and Riester, 2012) DIRNDL	old (n.a.)				79.61% ⁵	6,668 (10-fold cross-validation)
	mediated (n.a.)					
	new (n.a.) other (n.a.)					
(Markert et al., 2012) OntoNotes ⁶	old (29.48%)	88.6%	81.7%	85.0%	79.4%	10,980 (10-fold cross-validation)
	mediated (33.77%)	77.4%	68.4%	72.6%		
	new (36.75%)	75.1%	87.7%	80.9%		

∅ average values,

* not reported in the paper, but calculated based on the numbers reported.

1. 1 of 20 texts held out at a time for testing.
2. (all features/without span/neither span nor LSA)
3. These are the values as reported by Nissim. However, a re-calculation of accuracy as the weighted average of the reported precision values results in values of 84.9% (for two classes) and 78.8% (for three classes), respectively.
4. Distribution on training set. Test set distribution: 47.4% old, 36.6% med, 16.0% new.
5. For the 7-category distinction, accuracy is 75.82%, with a precision of 78.8%, recall of 75.5% and f-measure of 77.1% for the class *old*, having used automatic coreference detection as an input feature.
6. OntoNotes coreference annotation was extended to information status annotation according to Nissim et al. (2004). Distributions are calculated on the whole corpus unless stated otherwise; they may vary slightly across training/dev/test sets.

Table 4.5: State of the Art Classification Results (part II)

Ng and Cardie (2002) train a classifier to distinguish whether NPs are *anaphoric* or *not anaphoric* for application in a coreference resolution system. As a basis, they use the MUC-6 and MUC-7 data, training the classifiers on the ‘dry run’ data sets, and applying them to the ‘formal evaluation’ sets. No information is given on how the data is parsed and on the resulting number of NPs. Ng and Cardie use features from groups 3, 4, 5, 6, 10, and 11.

Uryupina (2003) trains a classifier for the category \pm *discourse_new*, with the goal of classifying *discourse_new* entities with high precision (in order to ‘lose’ as few candidates for coreference resolution as possible). As a basis, she uses the 20 texts from the Formaleval set of the MUC-7 corpus, parsed using Charniak’s (2000) parser. The data contains 3,710 noun phrases in total. Features from groups 3, 5, 10, 11 and 12 are used. The experiment was carried out as follows: one text at a time is held out. The remaining texts are used to first optimize Ripper classifier parameters (using 5-fold cross validation) and then train a model. The resulting model, in turn, is applied to the held-out text. The average performance on these texts is reported.

Hempelmann et al. (2005) model a three-way distinction between *given*, *new*, and *inferable* NPs. As basic data, they use four 4th grade textbook texts. These texts, containing 478 NPs, are manually annotated; inter-rater agreement is reported with $\kappa = .72$ for a Prince (1981)-based 5-way distinction (disagreement is reported for 18% of the cases). The class labels are encoded as a numeric scale for applying ordinal logistic regression, and conflated to a 3-way distinction (*given*, *new*, and *inferable*). The features used are from groups 5 and 11 (LSA and span represent measures of ‘indirect’ similarity with the context: pieces of texts are similar if their word cooccurrences are similar).

Nissim (2006) classifies *old* vs. *mediated* vs. *new* NPs and, among other things, reports on experiments where the classes *mediated* and *new* are conflated. As a basis, she uses the Switchboard corpus, splitting it into a training set (40,865 NPs), a development set (10,565 NPs) and an evaluation set (12,624 NPs), with instances randomized in a way that “NPs from the same dialogue were possibly split across the different sets” (Nissim, 2006, p. 95). The features she uses are of groups 1, 2, 5, 7 and 11. Nissim also reports on the contribution of each feature, evaluating single-feature and leave-one-out classifiers.

Denis and Baldridge (2007) classify *anaphoric* vs. *non-anaphoric* NPs using the ACE corpus²¹. They use features from groups 2, 5 and 11 as an input for a joint model of anaphoricity classification and coreference resolution formulated in ILP (Integer Linear Programming).

Ng (2009) trains a joint model for *anaphoricity* determination and coreference resolution using the ACE corpus. The features they employ are the same as in Ng and Cardie (2002) (this earlier study was based on the MUC corpora). Unfortunately, Ng does not provide an evaluation of the anaphoricity determination component in isolation, but only an evaluation of its effect on coreference resolution: he reports significant increases in MUC and CEAFF²² F-scores for all three types of news text.

Uryupina (2009), building on her earlier work (Uryupina, 2003), classifies NPs as \pm *discourse_new*. Again, she uses the MUC-7 corpus, parsed with Charniak’s (2000)

²¹As mentioned before, only certain types of entities are annotated for coreference in ACE, e.g. person, organisation etc.

²²MUC (Vilain et al., 1995) and CEAFF (Luo, 2005) are some of the most widely used scoring systems in anaphora and coreference resolution.

parser; this time using additional annotation of named entity types performed by the C&C NE-tagging system (Curran and Clark, 2003). The documents of the ‘Dryrun’ set are used for training, the ‘Formaleval’ set is used for the evaluation.

Rahman and Ng (2011) classify for *information status* like Nissim (2006), reusing her annotation of the Switchboard corpus. They randomly split the corpus into sets for training (nearly 88%) and testing (12%), maintaining the documents (unlike Nissim, who randomly split the instances). The feature set they use includes features from groups 1, 2, 4, 5, 7, 8 and 11.

Zhou and Kong (2011) classify for *anaphoricity*, using the ACE corpus. They use features from groups 4, 5, 7, 8, 10 and 11. Their label-propagation based approach with a polynomial kernel yields accuracies between 71.8% and 76.2%, depending on the domain (newswire vs. newspaper vs. broadcast news).

Cahill and Riestler (2012) classify for *information status* in 4 different granularities (one coarse-grained 4 category distinction comparable to Nissim’s (2006) scheme²³, and others with 7, 10 or 12 categories, respectively). The features they use are in groups 2, 4, 5, 6, 7, 8, 10 and 11, as well as automatically detected coreference information. Their 10-fold cross-validation experiment yields results for average accuracy between 69.56% (12 categories) and 79.61% (4 categories).²⁴

Markert et al. (2012) classify for *information status* similar to Nissim (2006) and Rahman and Ng (2011) (coarse-grained: 3 categories *old*, *mediated*, *new*, fine-grained: 9 categories, i.e. with six subtypes of *mediated*), using OntoNotes. Their features are included in groups 1, 2, 4, 5, 6, 7, 8 and 11. They use collective classification, performing 10-fold cross-validation. For comparison purposes, they reimplemented and applied Nissim’s (2006) and Rahman and Ng’s (2011) classifiers, which are outperformed by their model. To conclude this section, no comparisons are possible between any of the approaches presented above, mainly due to different categories and amounts of training data. As for different features, their impact is dependent on the definition of the categories, but also on the experiment settings (choice of training set) and on the features they are combined with. The impact of each feature is studied in Nissim (2006); the impact of different feature sets is studied in Hempelmann et al. (2005), Uryupina (2009), Rahman and Ng (2011), Cahill and Riestler (2012) and Markert et al. (2012).

To summarize the state of the art, there are two groups of classification approaches: those to discourse-givenness (with 2 possible values, *discourse-given* and *not discourse-given*) and those to information status (with more than 2 possible values). The former are based either on MUC or ACE corpora. The MUC corpora represent relatively small data sets, which do not include syntactic annotation (to the consequence that different studies use different NP boundaries). As a result, the models trained on this data are not comparable. The definition of coreference in the MUC annotation scheme has been criticized as being too lax (including, e.g., predications, appositions and function-value relations); that in the ACE corpus is limited to entities of certain predefined types (persons, organizations, locations etc.). The latter approaches are based on Switchboard, OntoNotes or DIRNDL. The works on Switchboard exclude non-referring expressions beforehand.

²³Nissim’s scheme conflates Riestler’s classes *given* and *situative* to form the class *old*.

²⁴When using the gold standard coreference annotation as additional clues, results range between 76.62% (12 categories) and 84.76% (7 categories). Here, only the best results are extracted; Cahill and Riestler’s study includes more experiments with different subsets of the feature set.

Chapter 5

Discourse-Givenness Classification: New Experiments and Results

This chapter provides information on the data and methods employed in this work’s experiments and motivates the choices taken. It also contains a presentation and discussion of the results.

5.1 Data

The corpora used for training the models are MUC-7, OntoNotes 1.0 and ARRAU (for English) and TüBa-D/Z 6.0 (for German).¹ Each corpus is used with its original annotations; the corresponding annotation schemes have been set into relation in Chapter 3. The resources used include three English corpora and one German corpus. I experiment with several corpora for one language (English) to avoid an overfitting of models to one annotation scheme. An additional corpus in another language (German) is employed to avoid an overfitting of models to special traits of one language, e.g. the relatively poor inflectional morphology and the writing of compounds as separate words in English. The following two examples illustrate linguistic differences between English and German relevant for resolving coreference. Features commonly used for discourse-givenness classification and coreference resolution in English include ‘*same head*’² and ‘*grammatical function*’. These features are popular because preprocessing tools for English are readily available and work in a simpler way than for many other languages, e.g. (i) lemmatization (determining a word’s uninflected form) and (ii) parsing (among other things, determining a constituent’s grammatical function). Inflection, for instance, in many cases involves the suffigation of *s* (for plural forms of nouns and for third person singular verbs). As to grammatical function, the subject is usually the constituent before the main verb. German, in contrast, has four inflectionally marked cases and relatively free constituent order. Preprocessing is thus more complex.

¹PCC requires a relatively large amount of background knowledge, either from articles in the same newspaper, or regional and time-specific knowledge, to interpret the commentaries. This is expected to complicate the classification task; experiments to that purpose remain to be done in future research.

²In anaphoricity classification, this is a feature that is true iff there is a noun phrase in the previous context that has the same head word as the current noun phrase. In coreference resolution, this feature is true for a pair of noun phrases that have the same word as their head word.

5.2 Methods

The goal of this work is to find methods for an optimal modelling of discourse-givenness of noun phrases in English and German. As the performance of a classification model depends on the discriminatory power and combinational interaction of the features it uses, one of the tasks is to find additional features that complement previously suggested features. The features newly introduced in this work mainly cover the area of semantic similarity and the use of the NPs' local context. All features used are described in Section 5.2.1.

Another influence of the model's performance is the learning algorithm and its aptness to the respective task. The decisions regarding algorithms are shortly discussed in Section 5.2.2.

In addition, classification can profit from taking into account some of the data's characteristics, e.g. its skew class distribution: the vast majority of NPs is not anaphoric (see Table 4.4; more details will be given in Sections 5.3.1.1 to 5.3.1.4). On that score, the choice of evaluation measures is crucial (see the discussion in Section 4.3.1).

In some of the previous work (Nissim (2006), Rahman and Ng (2011) and Markert et al. (2012)), the classifier is trained and applied only to a subset of the instances:³ expressions that are nonreferring, temporal, locational, directional etc. are excluded. What remains is, for the most part, instances of specific reference (e.g. reference to persons, organizations, etc.), with occasional instances of reference to kinds, abstract concepts and the like.⁴ This restriction leads to a more balanced class distribution, which a classifier can learn more easily. The aim of the present work, however, is to help a coreference resolution system performing on full text (i.e. any instance), not on a subset of the data, which for the time being would have to be selected manually.

5.2.1 Features

Each feature stands for a categorization or measurement of a certain aspect of a noun phrase or its relation to the context.⁵ A feature's usefulness in a model depends, among other things, on the coding of the features in the model, and on how the learning algorithm rates and/or combines them. For instance, Nissim (2006) experiments with an NP's head word as a feature (hoping that the model would learn some mediated entities) in decision trees, getting a negative effect on the model's performance. Rahman and Ng (2011), however, report on the successful use of this feature (in boolean coding) in their SVM model.

An overview of the features employed in my experiments is given in Table 5.2. Numbering of feature groups follows Sections 4.2.1 and 4.2.2. New features that have not been used in previous work are marked (*).

The newly introduced features are supposed to measure (i) how specific (or abstract, respectively) an expression is, (ii) how similar single words contained in the expression are to the words in the previous context, and (iii) whether the (local) context provides hints on the expression's discourse-givenness.

³Besides this, deictic personal pronouns are always categorized as *old*.

⁴This measure leads to a coreference definition close to that realized in the ACE corpus.

⁵Only NPs will be regarded; possessive pronouns, WHNPs etc. are not included in the experiments.

Abstractness/Specificity

News texts usually report on events involving concrete referents. Concrete referents tend to be described rather specifically, often resulting in longer expressions (*length_in_chars*). Also, embedded definite phrases or expressions with a possessive ‘s’ often have specific referents: typical possessors are e.g. persons or organizations. On the other hand, certain suffixes point to abstract common nouns, such as *-ion*, *-ness* etc. (*suffix_n*).

Specificity is a precondition for coreference in many corpora. I understand specificity as a feature representing two components:

- a) an expression’s descriptive content (preciseness of description), and
- b) relatedness to the context it occurs in.

When referring to a particular object, the speaker needs to distinguish it a) from other objects in the world, and b) from other discourse referents in the context.

A distinction from other objects in the world can be accomplished by using proper names or precise descriptions (e.g. *the blue book* or *the lexicon* is more precise than *the book*).

A distinction from other discourse referents can be made by using different concepts, or, when re-using lexical material, by adding descriptive content (e.g. *the town - the new town*, *a bus - another bus*) or by using the indefinite plural form (e.g. *Investigators continued their search. At Calverton, investigators began piecing together the airplane*).

For an operationalisation of specificity, I assume that

- 1) the more precise the description, the fewer documents it occurs in, and
- 2) the more important a discourse referent is for a discourse *d*, the more frequently it occurs in *d*.

As a measure for these two components of specificity, I use

- 1) inverse document frequency (idf) and
- 2) term frequency (tf), combined to tfidf, a well-known measure from the field of Information Retrieval.

For the purpose of discourse-givenness classification, tfidf is calculated using the full expression as a term on the one hand, and using a sliding window of characters as a term on the other hand (4 characters in the experiments presented here). The sliding window was implemented to make the method more robust against inflection, composition, and the shortening of names (e.g. *Alex* for *Alexander*), considering in particular that it is applied not only to English.

The calculation of tf-idf-related features is sketched at the end of this section.

Similarity

Synonyms are sometimes used for referring to the same entity for stylistic reasons, e.g. to avoid word repetitions. In earlier work, latent semantic analysis and variants (Hempelmann et al., 2005) have been used, as well as semantic relatedness, measured e.g. by means of WordNet or GermaNet (Markert et al., 2012; Cahill and Riester, 2012). In the present work, semantic similarity as calculated by DISCO (Kolb, 2008) is used in the feature *maxsimilar_mention*.⁶ The calculation of this feature is sketched at the end of this section.

Context

Motivated by the findings in the theoretical part (Section 2.3), I also introduce features exploiting an expression’s context. These contextual features are designed to complement

⁶Distributional similarity (LSA and span) between all words of the NP and the NP’s preceding context has been used by Hempelmann et al. (2005) in their logistic regression experiments.

the other features, e.g. similarity and identity. The syntactic node dominating the NP (feature *mother_cat*) can help discourse-givenness classification: an NP dominated by an adjective phrase (ADJP), for instance, is less likely to be *given* than an NP dominated by a prepositional phrase (PP).⁷ Information on the verb’s tense, change in tense compared to the previous sentence, and whether the verb has been used previously can also hint to *given* expressions (for instance, the verb’s subject). These features are operationalized as the verb’s suffix (*v_suffix*), a boolean value representing whether the verb’s suffix is equal to the preceding verb’s suffix (*equal_v_suffix*), and a numeric value (*v_previous*) representing the number of times the verb has been mentioned. The size of the expression’s left context (in tokens) is also taken into account (*size_left_context*), as the first few NPs in a text are rarely *given*.⁸

The impact of contextual features (group 8) is particularly interesting from the theoretical perspective. These features will play a prominent role in the classification experiments.

Calculation of tf-idf-related Features

This subsection summarizes the calculation of tf-idf-related features as described in Ritz (2010). For a term t in a document d , tfidf is commonly calculated according to the formula in (114).

$$(114) \quad \text{DEFINITION tfidf} \\ \text{tfidf}_{t,d} = \text{tf}_{t,d} * \text{idf}_t \text{ with } \text{tf}_{t,d} \text{ the relative frequency of } t \text{ in } d \text{ and} \\ \text{idf}_t = \log\left(\frac{|D|}{|D_t|}\right), \text{ with } D \text{ the document collection, } D_t \text{ the documents containing } t$$

For the feature ‘mention_tfidf’ (and ‘mention_idf’), tfidf (and idf) is calculated for the whole NP. For the features ‘sum_tfidf’, ‘max_tfidf’ etc., the NP is sliced into terms by a sliding window (here, a window of 4 characters is used). Across the tfidf values of each term, the sum and maximum are calculated (see 115 and 116, respectively).

$$(115) \quad \text{sum}_{\text{tfidf}_{NP_s^e, d_s}} = \sum_{i=s}^{e-l+1} \text{tfidf}_{t_i, d_i} \text{ with } l \text{ the window size (here: 4)}$$

$$(116) \quad \text{max}_{\text{tfidf}_{NP_s^e, d_s}} = \max_{i \in [s, e-l+1]} \text{tfidf}_{t_i, d_i}$$

Additionally to the term frequency, $\text{tf}_{t, \overline{d_k}}$, the increase in term frequency with the current mention, is calculated and used as a replacement for term frequency.

$$(117) \quad \text{tf}_{t, \overline{d_k}} = \text{tf}_{t,d} - \text{tf}_{t, d_k}, \text{ with } k \text{ the starting position of the current NP and } d_k \text{ document } d \text{ up to character position } k.$$

Sum and maximum are calculated as above. For all tf-idf-related features, the whole corpus is used as the document collection.

Calculation of maxsimilar_mention

The feature ‘maxsimilar_mention’ is calculated as follows: for each head word, the similarity between this word and each of its preceding words in the text is calculated using

⁷To a certain extent, the feature *n_pos_left/right* captures similar information, for instance, an expression dominated by a PP node is preceded by a preposition; the part of speech tag to its left would thus be IN (the tag for prepositions or subordinating conjunctions).

⁸This feature is related, but not equal to Ng and Cardie’s (2002) feature *position* with the values ‘*first sentence*’, ‘*first paragraph*’, ‘*header*’. It is a numeric feature. Typographic information (information on paragraphs and headers) is not available for most corpora.

DISCO (Kolb, 2008). Consider Example 118 (repeated from Example 16 above). The similarity values for this example are given in Table 5.1, maximum values are printed in bold face (‘n.a.’ means one of the words does not occur in the index, either because it occurred too infrequently in the source texts, like *Miami-based*, or too often, like *of, the*).

(118) Four former Cordis Corp.₁ officials₂ were acquitted of federal charges related to the Miami-based company’s₁ sale of pacemakers_{3,4,5}, including conspiracy to hide pacemaker defects_{6,7}.

	officials	charges	company	sale	pacemakers	conspiracy	defects
Four	0.0020	0.0095	0.0005	0.0022	0.0	0.0010	0.0008
former	0.0058	0.0008	0.0050	0.0024	0.0	0.0	0.0
Cordis	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
Corp	0.0031	0.0004	0.0142	0.0073	0.0039	0.0014	0.0
officials		0.0169	0.0066	0.0079	0.0	0.0187	0.0008
were		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
acquitted		0.0511	0.0	0.0	0.0	0.0313	0.0009
of		n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
federal		0.0238	0.0072	0.0135	0.0	0.0225	0.0009
charges			0.0075	0.0087	0.0	0.0684	0.0070
related			0.0018	0.0037	0.0	0.0031	0.0010
to			n.a.	n.a.	n.a.	n.a.	n.a.
the			n.a.	n.a.	n.a.	n.a.	n.a.
Miami-based			n.a.	n.a.	n.a.	n.a.	n.a.
company				0.0129	0.0013	0.0049	0.0025
s					n.a.	n.a.	n.a.
sale					0.0	0.0109	0.0030
of					n.a.	n.a.	n.a.
pacemakers						0.0	0.0097
including						0.0028	0.0009
conspiracy							0.0081
to							n.a.
hide							0.0046
pacemaker							0.0107
defects							

Table 5.1: Feature ‘maxsimilar.mention’: Calculation Example (column maximum in bold face)

Considered in isolation, the values of these features might not predict the discourse-givenness of an NP. Together with other features, however, e.g. the NP’s determiner, they help discriminate discourse-given NPs.

(1) Surface Form	
suffix_ <i>n</i> *	e.g. <i>ter</i> , <i>ist</i> , <i>ion</i> , ... for $n=3$
(categorical)	suffix (last n characters) of the NP's head lemma, where n is between 1 and 3
(2) Length	
length_in_chars*	
(numeric)	the NP's length in characters
length_in_tokens	
(numeric)	the NP's length in tokens
(3) Spelling	
all_capitalized	{1, 0}
(boolean)	1 if the NP consists only of capital letters and special characters (e.g. <i>ABC</i> , <i>A&M</i> , <i>U.S.</i>), else 0
(4) Morphological Features	
pron_morph	e.g. 1.pl, 2.sg/pl, 3.sg.f
(categorical)	if the NP is a pronoun: person, number and, in case of 3rd person, gender features of this pronoun in English, number of proper and common nouns is encoded in the feature 'n_form' values <i>NNS</i> and <i>NNPS</i> , respectively in German, all the morphological information available is used (number, person, gender of the head)
(5) Syntactic Form and Structure	
phrase_form	{def, indef, dem, pronposs, pron, pronrel, interrog, n.a.}
(categorical)	<i>def</i> if NP dominates a definite determiner, <i>indef</i> if it dominates an indefinite determiner, <i>dem</i> if it dominates a demonstrative determiner, <i>pronposs</i> if it dominates a possessive, <i>pron</i> if it consists of a personal pronoun, <i>pronrel</i> if it consists of a relative pronoun, <i>interrog</i> if it consists of an interrogative pronoun, else <i>n.a.</i>
DT_form	e.g. <i>the</i> , <i>both</i> , <i>many</i> , <i>such</i> , etc.
(categorical)	lemma of the NP's determiner if present
DT_type	{DT, WDT, PDT} for English, {ART, PIAT} for German
(categorical)	type of leftmost determiner dominated by the NP

	<p>for English: <i>DT</i> determiner, <i>PDT</i> predeterminer (<i>half/PDT the/DT level/NN</i>), <i>WDT</i> wh-determiner (<i>the show on which/WDT Ms. Chung appears</i>)</p> <p>for German: <i>ART</i> definite or indefinite determiner, <i>PIAT</i> attributive indefinite pronoun without determiner like in <i>kein/PIAT Mensch</i>)</p>
n_form (categorical)	<p>{NN, NNS, NNP, NNPS} for English, {NN, NE} for German</p> <p>pos tag of the NP's rightmost noun (if it contains a noun)</p> <p>for English: NN common noun, NNP proper noun, suffix S plural form</p> <p>for German: NN common noun, NE proper noun</p>
embedded_phrase_form* (categorical)	<p>{def, indef, dem, pronposs, pron, pronrel, interrog, n.a.}</p> <p>if the NP embeds another NP, the form (analogous to <i>phrase_form</i>) of this embedded NP</p>
possessive_s* (boolean)	<p>{1, 0}</p> <p>1 if the NP's rightmost token is an apostrophe or apostrophe+'s', respectively</p> <p>In TüBa-D/Z, morphological information of the head is used instead.</p>
(6) Semantic Class	
ne_type (categorical)	<p>e.g. <i>organization, person, location, product, language, etc.</i></p> <p>taken from the Penn Treebank annotations (in Penn Treebank, only Named Entities consisting of exactly one word are annotated with their respective Named Entity type). This feature is only available for OntoNotes and ARRAU.</p>
has_title (categorical)	<p>e.g. <i>Mr., Mrs., Dr., Rep., etc.</i></p> <p>if the NP contains a token matching the regular expression “$^{\wedge}[AZ][a-z]^{\wedge}.$”, this token represents the value of the feature</p>
(7) Grammatical Function and Position	
grammfunc (categorical)	<p>e.g. <i>SBJ, TPC, ADV, LOC, TMP, PRD etc.</i></p> <p>grammatical function taken from Penn Treebank annotation or TüBa-D/Z syntax annotation (<i>ON</i>: object nominative, i.e. subject, <i>OA</i> object accusative, <i>OPP</i> PP object, etc.)</p>
(8) Local Context	
mother_cat*	e.g. <i>NP, S, PP, VP, ADJP, etc.</i>

(categorical)	category of the phrase directly dominating the NP
<i>n_pos_left/right</i> (categorical)	e.g. ‘ <i>JJR NN TO</i> ’, ‘. <i>NNPS VBP</i> ’ for $n=3$, <i>left</i> combination of n pos tags to the left of the NP (or to the right, respectively) for $n \in \{1, 2, 3\}$
<i>v_previous*</i> (numeric)	number of times the clause’s verb has been mentioned before
<i>v_suffix_n*</i> (categorical)	suffix of length $n \in \{1, 2\}$ of verb of clause containing the current NP
<i>equal_v_suffix_n*</i> (boolean)	$\{1, 0\}$ true iff nearest preceding verb has the same suffix of length n ($n \in \{1, 2\}$)
(9) Saliency	
–	(not used in my experiments)
(10) Sentence Construction	
–	(predication is encoded via grammatical function value PRD)
(11) Agreement, Identity and Similarity	
<i>exact_previous_mention</i> (numeric)	number of times the NP has been mentioned before in the same text (in German, lemmatization using Treetagger (Schmid, 1994) is performed to cover mismatches which are only due to differences in case)
<i>head_previous_mention</i> (numeric)	number of times the NP’s head has been mentioned before in the same text (the lemma is used here)
<i>mention_tfidf/idf*</i> (numeric)	idf and tfidf with the precise NP form as the term
<i>sum/max</i> of ngram-based tf, ... * (numeric)	term frequency (tf), inverse document frequency (idf), and $tf*idf$ with 4grams of characters as the term; sum and maximum are calculated across all terms of an NP (Ritz, 2010)
<i>maxsimilar_mention*</i> (numeric)	maximum of similarity of the NP’s head word to any word in the previous context, calculated with DISCO (Kolb, 2008)

(12) Definiteness Probability	
–	(not used in my experiments)
(13) Other	
size_left_context (in tokens)* (numeric)	number of tokens before the NP

Table 5.2: Features Used in the Classification Experiments

5.2.2 Algorithms and Evaluation Measures

Algorithms used in previous work are described in Section 4.3. The experiments in this work will be carried out with a subset of these algorithms, in particular rule-based and similarity-based algorithms. Only recently, the question has been raised whether the classification of discourse-givenness could profit from the application of sequential models. This, however, is beyond the scope of this work.

The classification experiments are evaluated using selected standard measures, learning curves, comparisons to human interrater agreement (where reported), and comparisons to related work where applicable.

5.3 Classification Results

In this section, the results of the classification experiments are presented and interpreted. Effects of influence factors such as different features, algorithms and methods for splitting the data into sets for training, development and testing are investigated.

5.3.1 Quantitative Results

The splitting of the data into sets for training, development and testing is carried out randomly using WEKA’s supervised instance filter *Resample*, which produces samples with similar class distributions. Where available, original splits are also used in the experiments.

For each of the corpora, four sets of features are used: a *baseline* classifier, comparable to Nissim’s (2006)⁹, using the features *phrase_form*, *exact_previous_mention*, *head_previous_mention*, *grammfunc* and *length_in_tokens*, a classifier using *all* features presented in Section 5.2.1, a classifier ‘*no local context*’, which makes use of all features presented in Section 5.2.1 except those from group (8) and a classifier ‘*no new*’ which uses all features but those newly introduced (marked ‘*’ in Table 4.2). These two groups of features are held out to investigate the contribution of the newly introduced features and the role of the local context in the resolution of coreference.

Three different algorithms are used: decision trees (J48 in WEKA), Ripper (JRip in WEKA), and SVMs (SMO in WEKA, with a linear kernel), as these have been used successfully and most commonly in the literature. WEKA’s standard parameter settings are used.

Throughout the experiments, the results using J48 are slightly, in most cases significantly, better than those using JRip or SMO.

The evaluation of the classification results follows the methods explained in Section 4.3.1. The influence of the size of the training set is shown using learning curves (here, *f* measure of the smaller class *discourse-given* is used).

5.3.1.1 OntoNotes

OntoNotes is the largest of the English corpora, it will thus be used as a point of reference. Around 16% of NPs in OntoNotes are discourse-given (see the class distribution in Table 5.3).¹⁰

The corpus was randomly split into three sets, one for training, one for development, and one for testing. Documents were retained (i.e. instances from one article were not split across sets). Document length was controlled for in the random split to avoid a concentration of very short or very long documents in any of the sets. The distributions are shown in Table 5.4. Differences between sets are not significant (χ^2 test).

⁹Nissim’s features are (value sets are given in curly brackets, otherwise the value type is given in round brackets): full prev mention (numeric), mention time {first,second,more}, partial prev mention {yes,no,na}, determiner {bare,def,dem,undef,poss,na}, NP length (numeric), grammatical role {subject,subppass,object,pp,other}, NP type {pronoun,common,proper,other} (Nissim, 2006, p. 97).

¹⁰For determining an NP’s discourse-givenness, only the ident-relation is used; the appos-relation is disregarded.

+discourse-given	20,473	(15.78%)
-discourse-given	109,308	(84.22%)
total	129,781	(100.00%)

Table 5.3: OntoNotes 1.0: Class Distribution

Set	#NPs +discourse-given	(%)	#NPs -discourse-given	(%)	#NPs in total
Train	16,363	(15.79%)	87,271	(84.21%)	103,634
Dev	2,144	(15.98%)	11,270	(84.02%)	13,414
Test	1,966	(15.44%)	10,767	(84.56%)	12,733

Table 5.4: OntoNotes 1.0: Class Distribution in Training, Development and Test Set (random split)

Table 5.5¹¹ shows the J48 (decision tree)-based classifiers’ performance on the development set and on the test set, respectively.¹² The development set has been used for intermediate testing during the development process; results on this set are only reported to show that development set and test set are relatively consistent. Besides a random split, 5-fold cross-validation was carried out. In this setting, the instances from one document may be split across different sets. In the literature, either approach has been used, but reports on a direct comparison do not exist to my knowledge. The results are very similar to the results on the document-retaining split.

The results show a significant influence of the additional features on performance, in particular a substantial gain in recall.

Figure 5.1 gives an excerpt of the model (J48 tree, all features, trained on training set). Local context features can be found in several subtrees and at different depths in the decision tree.

As to the different learning algorithms tested, J48 performed best throughout the experiments. It performed slightly, in most cases significantly better than JRip and SMO (see Table 5.6).

¹¹Results of significance tests (McNemar with correction) are given in superscript: numbers refer to classifier numbers, the levels of significance are represented by asterisks (*). These levels are: ‘***’ highly significant ($p < 0.001$), ‘**’ very significant ($p < 0.01$), ‘*’ significant ($p < 0.05$). For instance, all^{1***,2*} means the performance of classifier *all* differs highly significantly from that of classifier 1 (baseline), and significantly from that of classifier 2.

¹²For a later comparison with MUC, which does not contain information on a constituent’s grammatical function, I ran additional experiments leaving out this feature, with the following results: For the classifier ‘*no local context*’, the results are not significantly different on the development set (Accuracy 92.11%, Precision 78.5%, Recall 69.6%, F-measure 73.8%) but very significant on the test set (Accuracy 91.83%, Precision 76.9%, Recall 67.3%, F-score 71.8%). For the classifier ‘*all*’, a leaving out of the feature ‘grammatical function’ does not make a significant difference, neither on the development set (Accuracy 92.83%, Precision 80.8%, Recall 72.2%, F-score 76.3%) nor on the test set (Accuracy 92.96%, Precision 80.8%, Recall 71.4%, F-score 75.8%).

training: Train, test: Dev	acc	P	R	F	class
1. baseline	89.39%	74.5%	51.1%	60.6%	+discourse-given
		91.2%	96.7%	93.9%	-discourse-given
2. no local context ^{1***}	92.38%	80.8%	68.7%	74.2%	+discourse-given
		94.2%	96.9%	95.5%	-discourse-given
3. no new ^{1***}	92.19%	80.2%	67.9%	73.6%	+discourse-given
		94.1%	96.8%	95.4%	-discourse-given
4. all ^{1***,2*,3***}	92.73%	80.1%	72.5%	76.1%	+discourse-given
		94.9%	96.6%	95.7%	-discourse-given
		<i>92.5%</i>	<i>92.7%</i>	<i>92.6%</i>	<i>weighted average</i>
training: Train, test: Test					
1. baseline	89.72%	76.6%	48.2%	59.1%	+discourse-given
		91.6%	97.7%	94.5%	-discourse-given
2. no local context ^{1***}	92.31%	79.5%	67.7%	73.1%	+discourse-given
		94.2%	96.8%	95.5%	-discourse-given
3. no new ^{1***}	92.26%	80.1%	66.4%	72.6%	+discourse-given
		94.0%	97.0%	95.5%	-discourse-given
4. all ^{1***,2**,3***}	93.02%	80.6%	72.1%	76.1%	+discourse-given
		95.0%	96.8%	95.9%	-discourse-given
		<i>92.8%</i>	<i>93.0%</i>	<i>92.9%</i>	<i>weighted average</i>
5-fold cross-validation on full dataset					
1. baseline	89.63%	75.0%	51.4%	61.0%	+discourse-given
		91.4%	96.8%	94.0%	-discourse-given
2. no local context	92.60%	80.0%	70.7%	75.1%	+discourse-given
		94.6%	96.7%	95.7%	-discourse-given
3. no new	92.43%	79.9%	69.5%	74.3%	+discourse-given
		94.4%	96.7%	95.6%	-discourse-given
4. all	93.11%	82.3%	71.7%	76.7%	+discourse-given
		94.8%	97.1%	96.0%	-discourse-given
		<i>92.9%</i>	<i>93.1%</i>	<i>92.9%</i>	<i>weighted average</i>

Table 5.5: OntoNotes 1.0: Classification Results (J48, random split/5-fold cross validation)


```

exact_previous_mention <= 0
| has_title = yes
| | max_tfreel <= 0.1182
| | | sum_idf <= 31.9596: NO (118.0/1.0)
| | | sum_idf > 31.9596
| | | | 1_pos_right = :: coref (0.0)
| | | | 1_pos_right = NNP: coref (0.0)
| | | | 1_pos_right = IN: coref (0.0)
| | | | 1_pos_right = CC: NO (1.0)
...
| | | | 1_pos_right = VBD: coref (10.0/3.0)
...
| | max_tfreel > 0.1182
| | | length_in_tokens <= 3
| | | | grammfunc = TMP-PRD: coref (0.0)
| | | | grammfunc = DIR: coref (0.0)
| | | | grammfunc = LOC-PRD: coref (0.0)
| | | | grammfunc = SBJ
| | | | | v_previous <= 20: coref (384.0/14.0)
| | | | | v_previous > 20
| | | | | | v_suffix_2 = 's: coref (0.0)
| | | | | | v_suffix_2 = to: coref (0.0)
| | | | | | v_suffix_2 = rs: coref (0.0)
...
| | | | | | v_suffix_2 = is
| | | | | | | 1_pos_right = :: NO (0.0)
| | | | | | | 1_pos_right = NNP: NO (0.0)
...
| | | | | | | 1_pos_right = VBZ
| | | | | | | equal_v_suffix_2 = 0: NO (19.0/3.0)
| | | | | | | equal_v_suffix_2 = 1: coref (2.0)
...
exact_previous_mention > 0
| n_form = NNP
| | mother_cat = S: coref (1165.42/30.7)
| | mother_cat = SQ: coref (6.18/1.0)
| | mother_cat = ADJP: NO (1.03/0.03)
...
| | mother_cat = NP
| | | 1_pos_right = :: NO (8.01/3.01)
| | | 1_pos_right = NNP: NO (4.01/0.01)
| | | 1_pos_right = IN
| | | | grammfunc = TMP-PRD: NO (0.0)
| | | | grammfunc = DIR: NO (0.0)
...

```

Figure 5.1: OntoNotes 1.0: Decision Tree (excerpt)

training: Train, test: Test	acc	P	R	F	class
4a. all (SMO)	91.84%	78.0%	68.1%	72.7%	+discourse-given
		94.1%	91.8%	91.6%	-discourse-given
		<i>91.5%</i>	<i>91.8%</i>	<i>91.6%</i>	<i>weighted average</i>
training: Train, test: Test					
4b. all (JRip) ^{4a***}	92.85%	83.3%	67.1%	74.3%	+discourse-given
		94.2%	97.5%	95.8%	-discourse-given
		<i>92.5%</i>	<i>92.8%</i>	<i>92.5%</i>	<i>weighted average</i>
training: Train, test: Test					
4c. all (J48) ^{4a***}	93.02%	80.6%	72.1%	76.1%	+discourse-given
		95.0%	96.8%	95.9%	-discourse-given
		<i>92.8%</i>	<i>93.0%</i>	<i>92.9%</i>	<i>weighted average</i>

Table 5.6: OntoNotes 1.0: Classification Results (SMO vs. JRip vs. J48, random split)

As for the influence of the number of training instances, the learning curve of classifier 4 (all features; evaluated on the randomly drawn sets for training and testing, respectively) is shown in Figure 5.2.

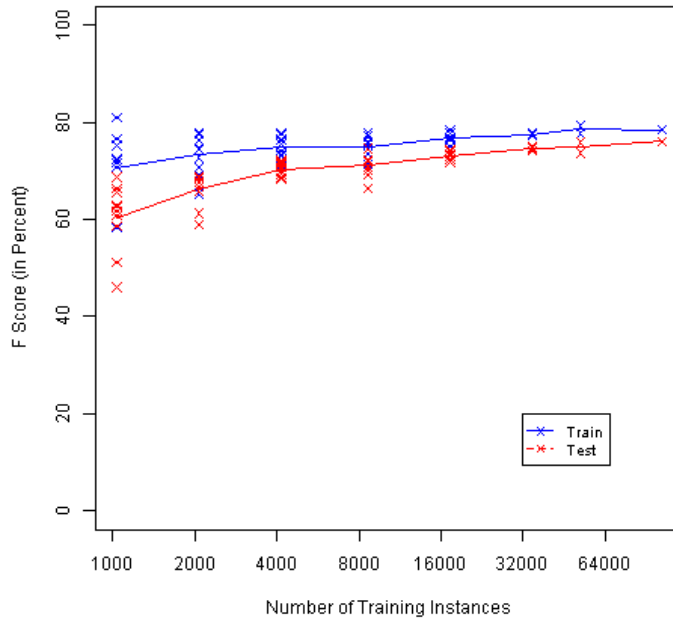


Figure 5.2: OntoNotes 1.0: Learning Curve

Each point represents the performance of a classifier trained on a randomly drawn subset of the size of its x value (note the logarithmic scale of the x axis). Lines represent means

values. Performance on the training set is drawn in blue. Testing the classifier on data it ‘has already seen’ in the training phase gives an upper bound of what can be expected, assuming that the data is consistently annotated throughout the sets. Performance on the test set is drawn in red. The curve shows a steady increase of F-measure with the number of training instances.

A detailed manual error analysis of the first 100 instances of classifier errors in the test set shows that the remaining classifier errors are of very different types. In particular, there are 6 annotation errors (see Example (119)) and 6 cases of nested annotations where the opening tags carrying the referent’s id are in the wrong order (see Example (120); errors marked with an asterisk ‘*’).

- (119) Limited volume ahead of the September trade data showed the market is nervous, but dealers added that the day’s modest gains also signaled some support₁ for London equities. They pegged *the support*_{*(missing: 1)} largely to anticipation that Britain’s current account imbalance can’t be much worse than the near record deficits seen in July and August.
- (120) In Tokyo, the Nikkei index₁ added 99.14 to 35585.52. [...] On Monday, traders noted that some investors took profits against the backdrop of the Nikkei’s₁ fast-paced recovery following its_{*2} plunge last Monday_{*1}. [...] Traders said the thin trading volume points to continued uncertainty by most investors following last Monday’s record 13% loss₂.

Some instances are especially hard to classify for the following reasons: 12 instances are aliases of a name mentioned previously, 7 contain additional information on the referent, and 4 instances are parts of idiomatic expressions (Example (121)). Another 4 instances occur in or relate to referents mentioned in direct speech, 3 refer back to referents in subheadings. 3 instances are presupposition anaphors (Example (122)), 2 are aggregations of referents mentioned previously, and 2 are cataphors.

- (121) NBC ’s Mr. Wright led *the way* in decrying the networks’ inability to match a Time-Warner combination.
- (122) Studios are “powerless” to get shows in prime-time lineups and keep them there long enough to go into lucrative rerun sales, he contends. And *that’s* why the rules, for the most part, must stay in place, he says.

The remaining 63 instances fit in neither of these classes; there is no obvious reason for their misclassification. It is striking, however, that 50 of the 100 inspected misclassified instances start with a definite determiner¹³, whereas this is the case for only 19% of NPs in the corpus in general. 12 of the inspected instances are pronouns, whereas in the corpus in general, less than 6% of NPs are pronouns.

As OntoNotes does not claim perfection – Hovy et al. (2006) is subtitled “the 90% solution” –, I manually inspected 600 randomly drawn instances. There were only 21 (3.5%) errors with respect to the discourse-givenness of the NP. Whether or not the NP was linked to the correct antecedent was disregarded. From this, we can estimate that the actual error rate for discourse-givenness in the corpus lies between 2% and 5%

¹³37 of them are +discourse-given, 13 of them -discourse-given. Most of the latter are related to the context via bridging.

(at 95% confidence).¹⁴ This is substantially better than the claim made for coreference annotation.

5.3.1.2 MUC-7

MUC-7 is about 10% the size of OntoNotes 1.0. The data does not contain syntax annotation, therefore Charniak’s (2000) parser was applied to it. Due to the parser’s output, the feature ‘grammatical function’ is not available for this data. During the mapping of coreference annotation to NPs, determiners outside the annotation span (see discussion in Section 3.2.4) were included in the span where necessary.

In the MUC-7 corpus, around 25% of NPs are discourse-given (see Table 5.7 for the class distribution).

+discourse-given	2,499	(25.08%)
-discourse-given	7,465	(74.92%)
total	9,963	(100.00 %)

Table 5.7: MUC-7: Class Distribution

As shared task data, MUC-7 is originally split into three sets: a training set (‘Train’), a set for a dryrun (‘Dryrun’), and a set for the formal evaluation (‘Formaleval’). Table 5.8 gives the more detailed class distributions in the respective sets. There are significant differences between the training and the other two sets with respect to class distribution.¹⁵

Set	#NPs +discourse-given	#NPs -discourse-given	#NPs in total
Train	156 (15.62%)	843 (84.38%)	999
Dryrun	1,401 (26.00%)	3987 (74.00%)	5,388
Formaleval	942 (26.33%)	2635 (73.67%)	3,577
Train+Dryrun	1,557 (24.38%)	4,830 (75.62%)	6,387

Table 5.8: MUC-7: Class Distribution in Train, Dryrun and Formaleval Set (original split)

In the first set of experiments, this original split was maintained; in a second set, the data is shuffled. The classification results of this first set of experiments are shown in Table 5.10. They include experiments with settings as used by Ng and Cardie (2002) and Uryupina (2009) for better comparability (trained on Dryrun, tested on Formaleval); their results are summarized in Section 4.4 and compared to this work in Section 5.4.2. Note, however, that the numbers of NPs, as well as the class distributions are not identical. This may be due to parsing and the mapping of coreference spans to noun phrases.

From the results, we observe that recall is remarkably low when training on the Training set only. The Training set is the smallest set. Besides that, it differs from the other sets, for instance with respect to the proportion of pronouns (see Table 5.9).¹⁶

¹⁴The interval is a binomial confidence interval; the test was carried out using R’s `binom.test`.

¹⁵Train vs. Dryrun ($\chi^2=49.32$, $p<0.0005$), Train vs. Formaleval ($\chi^2=49.20$, $p<0.005$).

¹⁶Differences between Train and Dryrun, and between Train and Formaleval are highly significant.

Set	pronouns	proportion of NPs
Train	118	11.81%
Dryrun	307	5.70%
Formaleval	204	5.70%

Table 5.9: MUC-7: Proportions of Pronouns in Different Sets

Figure 5.3 gives an excerpt of the classification model 4 (J48 decision tree, trained on 80% of Formaleval, using all features).

As for the effect of different feature sets, the picture is not clear: while in most settings, the addition of local context features does not seem to have a significant effect, it does in the fourth setting (training on Formaleval).

Due to the differences between sets, a second set of experiments was carried out with the data shuffled. The results are shown in Table 5.11. The results of the first two experiments, using 5-fold cross-validation, are comparable to the results on the original split regarding the effect of different feature sets; in particular, the quality of performance is comparable to that of the classifiers trained on the larger sets (Train+Dryrun or Formaleval or Dryrun, see Table 5.10). Also, the results of the two cross-validation settings are comparable, although in the second setting only 36% of the data was used.

Again, experiments with Uryupina’s settings have been carried out: an experiment using the Formaleval set only, holding out one document at a time for testing (Uryupina, 2003). JRip and SMO models perform slightly (in many cases significantly) lower, but similarly with respect to the dependence on the feature sets used.

As to the amount of training data, the classifier of course profits from additional training data from the Dryrun set. Particularly the Recall increases.

A learning curve of the classifier trained on Train and Dryrun, tested on Formaleval, is shown in Figure 5.4. It shows that a doubling of training instances from 3,200 to 6,400 does not increase the classifier’s performance. A learning curve of the shuffled data is shown in Figure 5.5. This curve shows an increase in performance up to 2,600 instances, and a slight decrease after that. This is probably due to differences between sets. Figure 5.6 again shows the curve based on the shuffled data, and added to this, the curve based on instances from the Formaleval set only (inserted in lighter colors). In comparison, the classifier trained on the Formaleval subset of the data constantly shows higher performance. This again is evidence that Formaleval is a highly consistent subset of the data. It is obvious that the evaluation set of a shared task should be, and is, most carefully annotated.

50 instances misclassified by a classifier using all features, trained on 4 of 5 folds (80%) of the Formaleval set were manually inspected. This error analysis reveals that the texts in MUC contain harder cases than in OntoNotes, in particular vague reference (in headlines), identity assertions and metonymy. 6 instances were annotation errors. 3 instances were parts of idiomatic or collocational phrases, 2 were expletives. MUC contains headlines, which give a very short summary of the text (e.g. *pilot dies in plane crash*). These often contain referring bare nouns, the referents of which are introduced properly only later.

```

exact_previous_mention <= 0
| mention_tfidf <= 0.1192
| | possessive_s = 1
| | | mention_tfidf <= 0.0269
| | | | DT_form = card.: NO (0.0)
| | | | DT_form = half: NO (0.0)
| | | | DT_form = the
| | | | | n_form = NNP
| | | | | | maxsimilar_mention <= 0.0311: NO (6.0/2.0)
| | | | | | maxsimilar_mention > 0.0311: coref (4.0)
| | | | | n_form = NNPS: NO (3.0)
| | | | | n_form = PRP$: coref (0.0)
| | | | | n_form = NN: coref (16.0/2.0)
| | | | | n_form = NNS: coref (3.0)
| | | | | n_form = na: coref (0.0)
...
| | possessive_s < 1
| | | phrase_form = indef: NO (651.0/56.0)
| | | phrase_form = def
| | | | prev_mention_time_tl <= 0
| | | | | v_previous_s_np <= 0: NO (1059.0/207.0)
| | | | | v_previous_s_np > 0: coref (67.0/31.0)
| | | | | prev_mention_time_tl > 0
| | | | | | mother_cat = S1: coref (0.0)
| | | | | | mother_cat = S: coref (72.37/15.68)
| | | | | | mother_cat = VP
| | | | | | | length_in_tokens <= 3
| | | | | | | | maxsimilar_mention <= 0.0041: NO (2.07/0.07)
| | | | | | | | maxsimilar_mention > 0.0041: coref (11.29/0.14)
| | | | | | | | length_in_tokens > 3: NO (2.14/0.14)
| | | | | | | mother_cat = SQ: coref (0.0)
| | | | | | | mother_cat = PRN: coref (0.0)
| | | | | | | mother_cat = ADJP: coref (0.0)
| | | | | | | mother_cat = NP: NO (57.89/17.35)
...

```

Figure 5.3: MUC-7: Decision Tree (excerpt)

training: Train, test: Dryrun	acc	P	R	F	class
1. baseline	75.98%	88.5%	8.8%	16.0%	+discourse-given
		75.7%	99.6%	86.0%	-discourse-given
2. no local context ^{1***}	79.08%	78.0%	27.3%	40.4%	+discourse-given
		79.2%	97.3%	87.3%	-discourse-given
3. no new ^{1***,2***}	75.30%	97.3%	5.1%	9.8%	+discourse-given
		75.0%	99.9%	85.7%	-discourse-given
4. all ^{1***,3***}	78.97%	83.2%	24.0%	37.2%	+discourse-given
		78.6%	98.3%	87.4%	-discourse-given
		79.8%	79.0%	74.3%	<i>weighted average</i>
training: Train, test: Formaleval					
1. baseline	75.17%	75.5%	8.5%	15.3%	+discourse-given
		75.2%	99.0%	85.5%	-discourse-given
2. no local context ^{1***}	74.48%	91.4%	3.4%	6.6%	+discourse-given
		74.3%	99.9%	85.2%	-discourse-given
3. no new ^{1***,2***}	79.06%	76.4%	29.6%	42.7%	+discourse-given
		79.4%	96.7%	87.2%	-discourse-given
4. all ^{1***,3***}	79.37%	85.7%	26.0%	39.9%	+discourse-given
		78.8%	98.4%	87.5%	-discourse-given
		80.6%	79.4%	75.0%	<i>weighted average</i>
training: Train+Dryrun, test: Formaleval					
1. baseline	82.89%	79.2%	47.6%	59.4%	+discourse-given
		83.6%	95.5%	89.2%	-discourse-given
2. no local context ^{1***}	84.88%	79.6%	57.2%	66.6%	+discourse-given
		86.1%	94.8%	90.2%	-discourse-given
3. no new ^{1**,2*}	83.95%	84.3%	48.0%	61.2%	+discourse-given
		83.9%	96.8%	89.9%	-discourse-given
4. all ^{1***}	84.54%	80.9%	54.0%	64.8%	+discourse-given
		85.3%	95.4%	90.1%	-discourse-given
		84.2%	84.5%	83.4%	<i>weighted average</i>
training: Formaleval, test: Train+Dryrun					
1. baseline	82.84%	72.1%	48.3%	57.8%	+discourse-given
		84.9%	94.0%	89.2%	-discourse-given
2. no local context	83.10%	67.6%	58.9%	63.0%	+discourse-given
		87.3%	90.9%	89.1%	-discourse-given
3. no new	83.09%	67.9%	58.1%	62.6%	+discourse-given
		87.1%	91.1%	89.1%	-discourse-given
4. all ^{1**,2*}	83.82%	68.7%	61.7%	65.0%	+discourse-given
		88.1%	91.0%	89.5%	-discourse-given
		83.3%	83.8%	83.5%	<i>weighted average</i>

Table 5.10: MUC-7: Classification Results (original split), part I

training: Dryrun, test: Formaleval (Ng and Cardie’s 2002, Uryupina’s 2009 setting)

1. baseline	82.97%	79.3%	47.9%	59.7%	+discourse-given
		85.9%	92.7%	89.2%	-discourse-given
2. no local context ^{1***}	84.99%	78.8%	58.8%	67.4%	+discourse-given
		86.5%	94.3%	90.3%	-discourse-given
3. no new ^{2**}	83.90%	82.3%	49.5%	61.8%	+discourse-given
		84.2%	96.2%	89.8%	-discourse-given
4. all ^{1***}	84.96%	81.2%	55.8%	66.2%	+discourse-given
		85.8%	95.4%	90.3%	-discourse-given
		84.6%	85.0%	84.0%	<i>weighted average</i>

Table 5.10: MUC-7: Classification Results (original split), part II

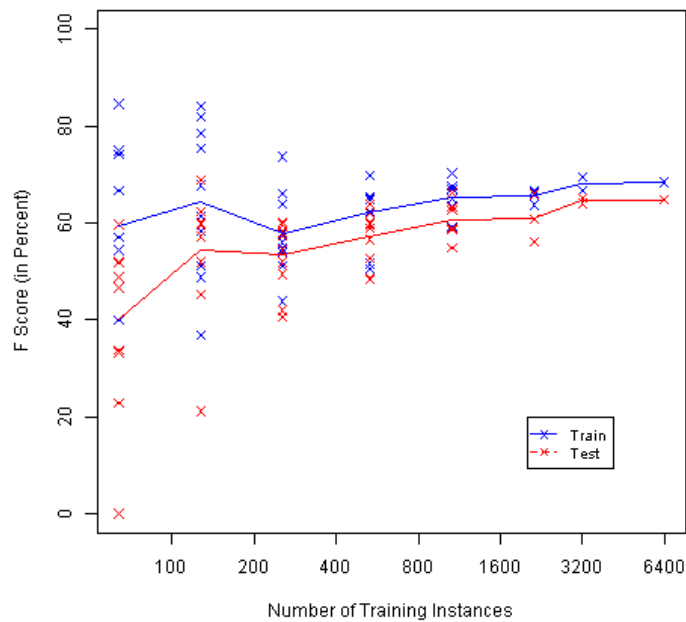


Figure 5.4: MUC-7: Learning Curve (original split, training: Train+Dryrun, test: Formaleval)

	acc	P	R	F	class
5-fold cross-validation using Train+Dryrun+Formaleval					
1. baseline	82.92%	74.8%	48.0%	58.5%	+discourse-given
		84.5%	94.6%	89.2%	-discourse-given
2. no local context	85.27%	78.3%	57.1%	66.0%	+discourse-given
		86.8%	94.7%	90.6%	-discourse-given
3. no new	84.32%	77.9%	52.3%	62.6%	+discourse-given
		85.6%	95.0%	90.1%	-discourse-given
4. all	85.33%	80.2%	55.0%	65.3%	+discourse-given
		86.4%	95.5%	90.7%	-discourse-given
		84.8%	85.3%	84.3%	<i>weighted average</i>
5-fold cross-validation using Formaleval					
1. baseline	82.94%	75.0%	47.9%	58.5%	+discourse-given
		84.5%	94.7%	89.3%	-discourse-given
2. no local context	85.27%	78.8%	56.4%	65.8%	+discourse-given
		86.7%	94.9%	90.6%	-discourse-given
3. no new	83.39%	74.7%	55.8%	63.9%	+discourse-given
		85.5%	93.2%	89.2%	-discourse-given
4. all	85.44%	79.7%	56.2%	65.9%	+discourse-given
		86.7%	95.2%	90.7%	-discourse-given
		84.9%	85.4%	84.5%	<i>weighted average</i>
hold out one document at a time using Formaleval (Uryupina's 2003 setting)					
1. baseline	83.04%	77.8%	47.1%	58.0%	+discourse-given
		83.5%	95.8%	89.2%	-discourse-given
2. no local context	83.92%	74.6%	55.7%	58.1%	+discourse-given
		85.6%	93.9%	89.5%	-discourse-given
3. no new	83.59%	72.7%	56.4%	62.7%	+discourse-given
		85.9%	93.1%	89.3%	-discourse-given
4. all	84.68%	76.0%	59.1%	66.0%	+discourse-given
		86.5%	93.9%	89.9%	-discourse-given
		84.4%	84.7%	83.9%	<i>weighted average</i>

Table 5.11: MUC-7: Classification Results (5-fold cross-validation, hold-out)

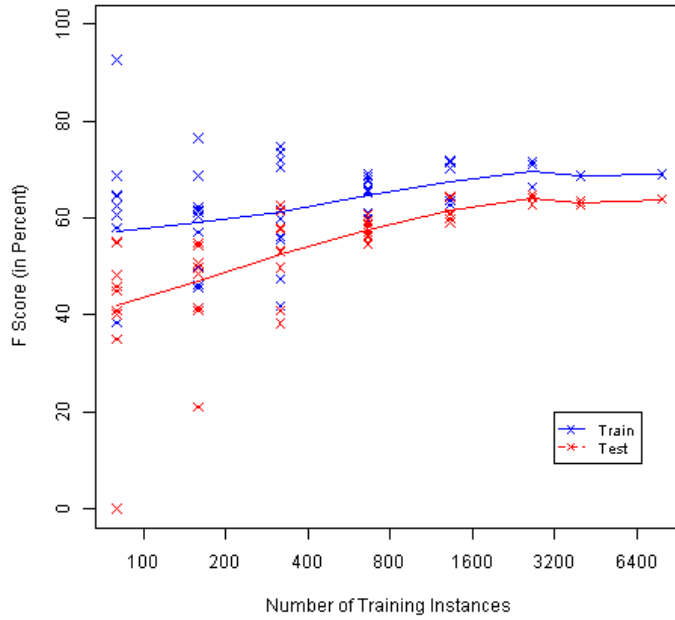


Figure 5.5: MUC-7: Learning Curve (random split)

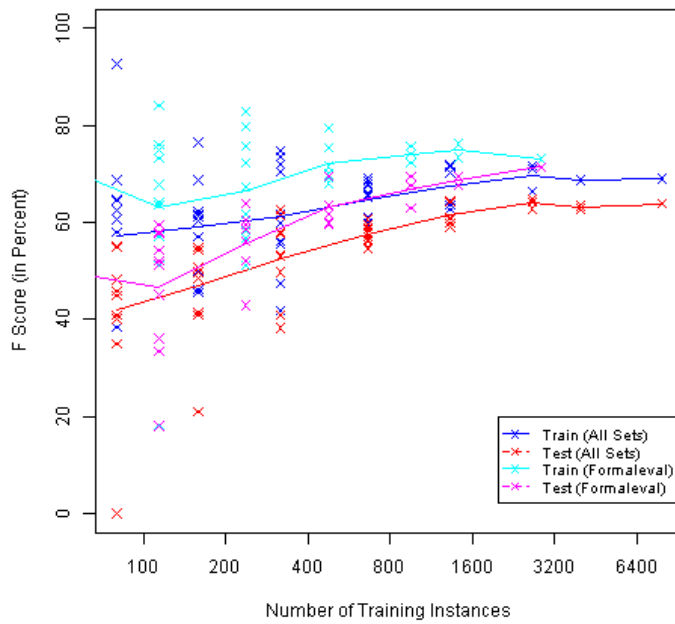


Figure 5.6: MUC-7: Learning Curve (random split; all instances vs. Formaleval)

In the error analysis, 2 cases could not be resolved to be referring back to a part of the headline. Additional challenges are predications (6 instances), entity-attribute relations (1 instance) and function-value relations (1 instance). Unfortunately and against the expectation, the local context features did not in all cases help in the classification of these phenomena. This is probably due to their relatively rare occurrence. Further, there was one case of a presupposition anaphor and one case of metonymy.

Similarly as in OntoNotes, 21 of 50 (42%) of the inspected misclassified instances start with a definite determiner, whereas in general (in the Formaleval set), this is the case with 27% of the instances. 6 of 50 (12%) are pronouns, whereas in general, the proportion of pronouns is less than 6%.

5.3.1.3 ARRAU

The ARRAU corpus, version 1.2, is about half the size of OntoNotes, and there is an overlap of 89 documents (texts from the *Wall Street Journal*), which contain 25,500 NPs (72,593 tokens). 23% of all NPs in ARRAU are discourse-given (see Table 5.12).

+discourse-given	14,381	(23.12%)
-discourse-given	47,828	(76.88%)
total	62,209	(100.00%)

Table 5.12: ARRAU 1.2: Class Distribution

The corpus was randomly split into sets for training, development and testing, see Table 5.13. There are no significant differences in class distributions between the sets.

Set	#NPs +discourse-given	(23.12%)	#NPs -discourse-given	(76.88%)	#NPs in total
Train	11,505	(23.12%)	38,262	(76.88%)	49,767
Dev	1,438	(23.12%)	4,783	(76.88%)	6,221
Test	1,438	(23.12%)	4,783	(76.88%)	6,221

Table 5.13: ARRAU 1.2: Class Distribution in Training, Development and Test Set (random split)

The classification results of J48 are given in Table 5.14. The baseline classifier performs similarly (measured in f-measure) as on OntoNotes, though ARRAU is only half the size of OntoNotes. The additional features, however, do not have a similarly large effect. Whereas the *newly introduced* features significantly improve the classification, the impact of *local context* features is positive on the development set but negative on the test set. The learning curve for the classifier using all NPs and all features is shown in Figure 5.7. Classification model 4 is shown in Figure 5.8.

50 instances misclassified by classifier 4 were manually inspected. This corpus differs from the other corpora in that it contains transcribed speech. This speech data contains a larger proportion of reference to situationally given objects (e.g. *the oranges*, *the boxcar* in the TRAINS dialogues), as well as some vagueness, e.g. *And it ends...*, where *it* could either refer to the story (previously mentioned in *It starts out there ...*) or it could be

training: Train, test: Dev	acc	P	R	F	class
1. baseline	82.90%	66.1%	53.3%	59.0%	+discourse-given
		86.7%	91.8%	89.2%	-discourse-given
2. no local context ^{1***}	87.88%	75.0%	63.6%	68.8%	+discourse-given
		89.5%	93.6%	91.5%	-discourse-given
3. no new ^{1***,2**}	86.69%	75.2%	60.4%	67.0%	+discourse-given
		88.8%	94.0%	91.3%	-discourse-given
4. all ^{1***,3***}	88.07%	81.0%	63.3%	71.0%	+discourse-given
		89.6%	95.5%	92.5%	-discourse-given
		87.6%	88.1%	87.5%	<i>weighted average</i>
training: Train, test: Test					
1. baseline	83.06%	66.2%	54.5%	59.8%	+discourse-given
		87.0%	91.6%	89.3%	-discourse-given
2. no local context ^{1***}	87.72%	75.8%	68.8%	72.1%	+discourse-given
		90.9%	93.4%	92.1%	-discourse-given
3. no new ^{1***,2**}	86.63%	75.5%	63.7%	69.1%	+discourse-given
		89.6%	93.8%	91.6%	-discourse-given
4. all ^{1***,3***}	87.59%	80.8%	60.7%	69.3%	+discourse-given
		89.0%	95.7%	92.2%	-discourse-given
		87.1%	87.6%	86.9%	<i>weighted average</i>

Table 5.14: ARRAU 1.2: Classification Results (random split, all NPs vs. referring NPs).

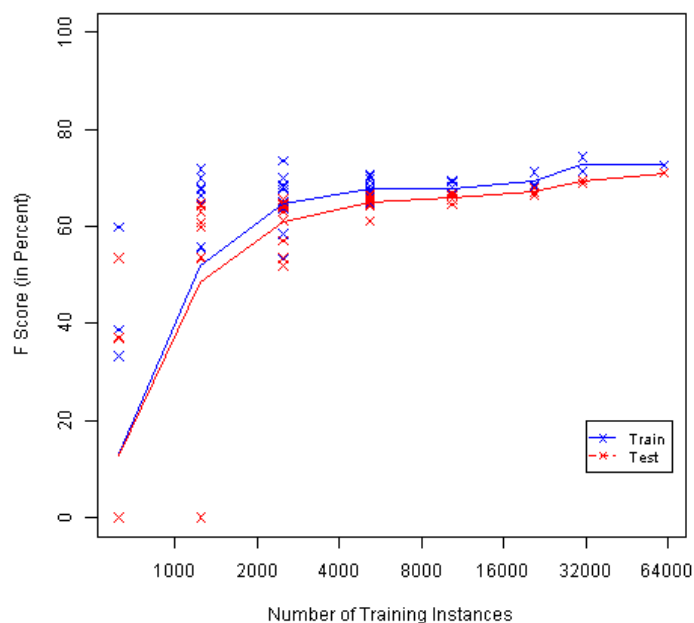


Figure 5.7: ARRAU 1.2: Learning Curve

```

head_previous_mention <= 0
| length_in_chars <= 4
| | phrase_form = def: NO (547.0/128.0)
| | phrase_form = pronposs: coref (4.0)
| | phrase_form = interrog: NO (2.0)
| | phrase_form = pron
...
| | | np_suffix_1 = I
| | | | 2_pos_left = VBN IN: coref (0.0)
| | | | 2_pos_left = NNP --: coref (0.0)
| | | | 2_pos_left = PRP VB: NO (2.01)
... | | | | 2_pos_left = PRP VBP
| | | | size_left_context <= 270: NO (2.02)
| | | | size_left_context > 270: coref (4.0)
...
| length_in_chars > 4
| | phrase_form = def: NO (5993.0/600.0)
| | phrase_form = pronposs: NO (668.0/49.0)
| | phrase_form = interrog: NO (2.0)
| | phrase_form = pron
| | | length_in_tokens <= 1: coref (67.0/10.0)
| | | length_in_tokens > 1: NO (23.0/6.0)
| | phrase_form = na: NO (19348.0/823.0)
head_previous_mention > 0
| all_capitalized = yes
| | ne_type = location: coref (2.0)
| | ne_type = person: NO (2.0/1.0)
| | ne_type = organization: coref (97.0/15.0)
| | ne_type = na: NO (1287.0/21.0)
| all_capitalized = NO
| | ne_type = location
| | | max_tf <= 508
| | | | maxsimilar_mention <= 0.0025: NO (136.0/45.0)
| | | | maxsimilar_mention > 0.0025: coref (423.0/96.0)
| | | max_tf > 508: coref (107.0/2.0)
...
| | | DT_form = both
| | | | sum_tfreel <= 0.2382: coref (14.0/1.0)
| | | | sum_tfreel > 0.2382: NO (11.0/1.0)
| | | DT_form = that
| | | | maxsimilar_mention <= 0.0566: coref (50.0/13.0)
| | | | maxsimilar_mention > 0.0566
| | | | | size_left_context <= 303: coref (4.0/1.0)
| | | | | size_left_context > 303: NO (17.0)
...

```

Figure 5.8: ARRAU 1.2: Decision Tree (excerpt)

analyzed as a kind of expletive. The referent is unclear in 4 cases. There are 2 annotation errors. 3 expressions are event anaphors, 3 are generic, and another 3 contain additional information on a previously mentioned referent. 2 are temporal expressions; one instance occurred in direct speech, one was contained in an apposition (according to the annotation scheme, only maximal phrases are annotated) and one instance was an idiomatic phrase. The other 30 could not be categorized.

There was one interesting case where two indefinite singular expressions refer to the same object (see Example (123); this example is from the Pear Stories section of ARRAU). In this case, identity of extension definitely holds, and, in my opinion, it also makes sense to classify this as coreference.

- (123) [...] Then a kid came along on a bicycle and parked the bicycle and checked the tree to make sure the man wasn't looking and he was about to steal a pear but he t he he picked up a basket Instead and went riding along the road on a bicycle And he passed a little girl on a bicycle and turned head and hit a rock and fell over and spilled the pears all over the road. In the meantime some other little kids came along and helped him pick up the pears and brushed him off and that sort of thing and he went on walking and one of them stopped him cause he had forgotten hat And took him hat and he gave them um he gave the three kids each a pear They were walking back in the direction uh toward the man who was picking pears in the pear tree and about the time he came down the pear tree and discovered that a basket of pears was missing

5.3.1.4 TüBa-D/Z

In TüBa-D/Z, relations to the context are labeled, allowing for a fine-grained distinction of *coreferential*, *anaphoric*, *bound*, *cataphoric*, *split_antecedent*, and *instance*. I experimented with two settings: in one, an NP is considered *+discourse-given* if it is related to the context via one of the following relations: *coreferential*, *anaphoric*, or *bound*. Otherwise, it is considered *-discourse-given*. According to this definition, 15% of the NPs are discourse-given. This setting is used for a rough comparison with the experiments presented above, in particular with OntoNotes, as both OntoNotes and TüBa-D/Z (but not MUC-7 and ARRAU) have specificity as a precondition. In the second setting, an NP is considered *+discourse-given* only if it is related to the context via the *coreferential* relation. This holds for less than 9% of the NPs (see Table 5.15 for the class distributions). The second setting corresponds more closely to the strict theoretical definition given in Section 2.3, but is also a more ambitious task.

<i>coreferential</i>	33,324	(8.91%)
<i>anaphoric</i>	21,405	(5.72%)
<i>bound</i>	1,304	(0.35%)
+discourse-given	56,033	(14.98%)
-discourse-given	317,902	(85.02%)
total	373,935	(100.00%)

Table 5.15: TüBa-D/Z 6.0: Class Distribution

The corpus was randomly split into three sets as shown in Table 5.16.¹⁷ There are no significant differences between the respective sets.

Set	#NPs +discourse-given	(14.98%)	#NPs -discourse-given	(85.02%)	#NPs in total
Train	33,620	(14.98%)	190,741	(85.02%)	224,361
Dev	11,207	(14.99%)	63,580	(85.01%)	74,787
Test	11,206	(14.98%)	63,581	(85.02%)	74,787

Table 5.16: TüBa-D/Z 6.0: Class Distribution in Training, Development and Test Set (random split)

Again, J48 performed best. Results of the J48 classifiers are shown in Table 5.17. Whereas the impact of the *newly introduced* features is significant, the impact of the *local context* features is not.

Compared to OntoNotes, the results of the *baseline* classifier are similar as on OntoNotes (around just above 60% F-measure), though the training set is approximately twice as big. The results of the classifier using *all features* are lower (around 73% vs. 76% on OntoNotes). This may be due to linguistic differences between English and German, in particular differences in the construction of compound nouns. In English, compound nouns are written separately, whereas in German, they are written in one word. See Example (124a) from OntoNotes (and its translation to German in b) for a coreferent instance using the same head noun as its compound antecedent but without the modifier. In English, this would be covered for by the feature *head_previous_mention*.

- (124) a. Hewlett-Packard Co. will announce today a software *program*₁ that allows computers in a network to speed up computing tasks by sending the tasks to each other. Called Task Broker, the *program*₁ acts something like an auctioneer among a group of computers wired together.
- b. Hewlett-Packard Co. wird heute ein *Softwareprogramm*₁ vorstellen, das es Computern in einem Netzwerk erlaubt, Rechenprozesse zu beschleunigen, indem sie sich diese gegenseitig zuschicken. Das *Programm* mit dem Namen Task Broker₁ agiert dabei wie eine Art Auktionator in einer Gruppe vernetzter Computer.

To a certain extent, this difference can be covered by the semantic similarity component. It is not fully compensated, however, as a good coverage of the vocabulary is harder to obtain for German, due to combinatorial explosion.

Results of the second setting for the grouping of categories (*coreferential* vs. all other) are shown in the second half of Table 5.17. This classification task yields lower results in general, as it is a harder task. Here, each feature group has a significant impact.

The learning curve is shown in Figure 5.9. It is worth noting that the difference between the performance on the training and that on the test set is consistently small, smaller than for any other corpus in this study.

An excerpt of the model learnt using all features for the first grouping of categories is shown in Figure 5.10, a corresponding model for the second grouping in Figure 5.11.

¹⁷This corpus is the largest and was split into sets with the proportions 60%-20%-20%.

	acc	P	R	F	class
trained on Train, tested on Dev (coreferential/anaphoric/bound vs. all other)					
1. baseline	89.73%	71.5%	52.4%	60.5%	+discourse-given
		92.0%	96.3%	94.1%	-discourse-given
2. no local context ^{1***}	92.90%	89.4%	59.7%	71.6%	+discourse-given
		93.3%	98.8%	95.9%	-discourse-given
3. no new ^{1***}	92.93%	88.6%	60.6%	72.0%	+discourse-given
		93.4%	98.6%	96.0%	-discourse-given
4. all ^{1***}	92.95%	87.6%	61.7%	72.4%	+discourse-given
		93.6%	98.5%	96.0%	-discourse-given
		<i>92.7%</i>	<i>92.9%</i>	<i>92.4%</i>	<i>weighted average</i>
trained on Train, tested on Test (coreferential/anaphoric/bound vs. all other)					
1. baseline	89.93%	72.1%	53.5%	61.4%	+discourse-given
		92.2%	96.4%	94.2%	-discourse-given
2. no local context ^{1***}	92.88%	87.1%	61.6%	72.2%	+discourse-given
		93.6%	98.4%	95.9%	-discourse-given
3. no new ^{1***,2**}	93.02%	89.0%	61.0%	72.0%	+discourse-given
		93.5%	98.7%	96.0%	-discourse-given
4. all ^{1***,3**}	93.14%	87.9%	62.9%	73.3%	+discourse-given
		93.8%	98.5%	96.1%	-discourse-given
		<i>92.9%</i>	<i>93.1%</i>	<i>92.7%</i>	<i>weighted average</i>
trained on Train, tested on Dev (coreferential vs. all other)					
1. baseline	92.40%	64.8%	32.2%	43.0%	+discourse-given
		93.7%	98.3%	95.9%	-discourse-given
2. no local context ^{1***}	94.20%	80.8%	45.9%	58.5%	+discourse-given
		94.9%	98.9%	96.9%	-discourse-given
3. no new ^{1***}	94.19%	83.8%	43.1%	56.9%	+discourse-given
		94.7%	99.2%	96.9%	-discourse-given
4. all ^{1***,2**,3***}	94.32%	82.6%	46.0%	59.1%	+discourse-given
		94.9%	99.1%	96.9%	-discourse-given
		<i>93.8%</i>	<i>94.3%</i>	<i>93.6%</i>	<i>weighted average</i>
trained on Train, tested on Test (coreferential vs. all other)					
1. baseline	92.55%	66.1%	33.6%	44.6%	+discourse-given
		93.8%	98.3%	96.0%	-discourse-given
2. no local context ^{1***}	94.35%	81.4%	47.4%	59.9%	+discourse-given
		95.1%	98.9%	97.0%	-discourse-given
3. no new ^{1***}	94.33%	84.8%	44.3%	58.2%	+discourse-given
		94.8%	99.2%	97.0%	-discourse-given
4. all ^{1***,2**,3***}	94.48%	83.8%	47.2%	60.4%	+discourse-given
		95.0%	99.1%	97.0%	-discourse-given
		<i>94.0%</i>	<i>94.5%</i>	<i>93.8%</i>	<i>weighted average</i>

Table 5.17: TüBa-D/Z 6.0: Classification Results (random split, *coreferential/anaphoric/bound* and *coreferential* vs. all other categories)

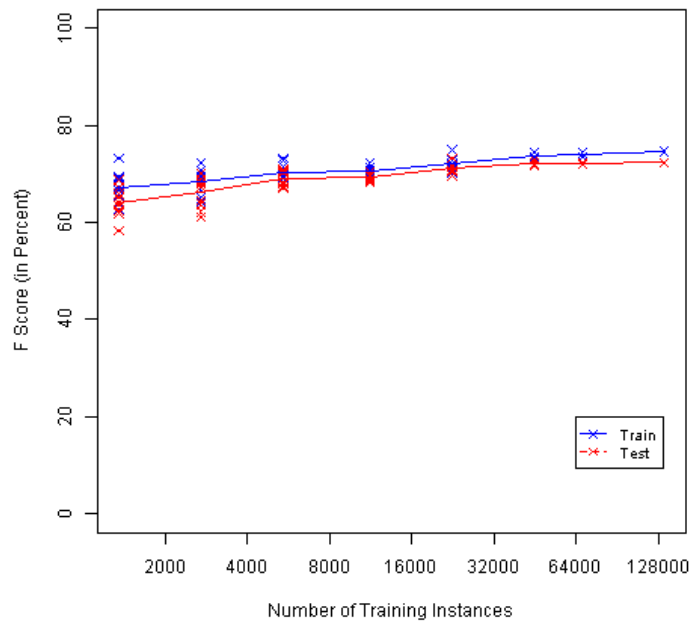


Figure 5.9: TüBa-D/Z 6.0: Learning Curve

100 of the misclassified instances of the classifier *coreferential/anaphoric/bound* vs. other categories were inspected manually. There were 9 annotation errors (missing coreference relations). 9 instances contained additional information to their antecedent. 6 have an entity-attribute relation to their antecedent, 4 expressions consist of a holonym of their antecedent expression, 2 are synonyms. 3 are short forms of named entities. 3 are reflexives. Among the errors, there is an overproportionally large amount of reflexive pronouns, demonstrative expressions, definite expressions and pronouns.

```

exact_previous_mention <= 0
...
| max_idf > -1
| | phrase_form = indef: NO (5966.89/79.0)
| | phrase_form = def
| | | maxsimilar_mention <= 0.0564: NO (39591.7/4041.19)
| | | maxsimilar_mention > 0.0564
| | | | grammfunc = ONK: NO (0.0)
| | | | grammfunc = PRED: NO (7.36/2.08)
...
| | | | grammfunc = APP: NO (199.98/0.24)
...
| | phrase_form = refl: NO (1058.44/172.86)
| | phrase_form = interrog: NO (163.72/1.95)
| | phrase_form = pronrel: NO (250.63/86.67)
| | phrase_form = indefpron: NO (4112.27/40.0)
| | phrase_form = pron
| | | morphology = asf: coref (0.0)
| | | morphology = ns*2: NO (0.95)
| | | morphology = d**: coref (0.0)
| | | morphology = as*1: NO (29.0/8.0)
...
| | | | | func = APP: NO (59.0/3.0)
...
exact_previous_mention > 0
| n_form = PRELAT: coref (0.0)
| n_form = PRELS: coref (3514.0/38.0)
...
| n_form = NN
| | phrase_form = indef: NO (154.35/10.0)
| | phrase_form = def
| | | grammfunc = ONK: NO (0.0)
| | | grammfunc = PRED: NO (6.0/1.0)
...
| n_form = PPER
| | morphology = asf: coref (0.0)
| | morphology = ns*2
| | | size_left_context <= 17: NO (10.0)
| | | size_left_context > 17: coref (63.0/10.0)
| | morphology = d**: coref (0.0)
...

```

Figure 5.10: TüBa-D/Z 6.0: Decision Tree (excerpt; *coreferential/anaphoric/bound* vs. all other categories)

```

contains_ne = yes
| exact_previous_mention <= 0
| | head_previous_mention <= 0: NO (28285.02/1735.0)
| | head_previous_mention > 0
| | | sum_idf <= 87.2244
| | | | n_form = PRELAT: coref (0.0)
| | | | n_form = PRELS: coref (0.0)
| | | | n_form = PTKZU: coref (0.0)
| | | | n_form = PTKNEG: coref (0.0)
| | | | n_form = PIAT: coref (0.0)
| | | | n_form = NE
| | | | func = ONK: coref (0.0)
| | | | func = PRED: NO (6.0/1.0)
| | | | func = OADVP-MO: coref (0.0)
| | | | func = PREDMODK: coref (0.0)
| | | | func = FOPP-MOD: coref (0.0)
| | | | func = OD-MOD: coref (0.0)
| | | | func = APP: NO (302.74/6.0)
...
| exact_previous_mention > 0
| | func = ONK: coref (0.0)
| | func = PRED: coref (16.16/4.16)
| | func = OADVP-MO: coref (0.0)
| | func = PREDMODK: coref (0.0)
| | func = FOPP-MOD: coref (0.0)
| | func = OD-MOD: coref (0.0)
| | func = APP: NO (702.65/30.0)
...
| | func = OD: coref (90.0/1.0)
| | func = FOPP: coref (0.0)
| | func = KONJ: coref (668.16/132.16)
| | func = HD
| | | 1_pos_left = $(: coref (119.0/58.0)
| | | 1_pos_left = VMFIN: coref (2.0/1.0)
...
contains_ne = no
| DT_form = der
| | exact_previous_mention <= 0
| | | func = ONK: NO (0.0)
| | | func = PRED
| | | | maxsimilar_mention <= 0.038: NO (67.64/1.0)
| | | | maxsimilar_mention > 0.038: coref (4.36/1.36)
...

```

Figure 5.11: TüBa-D/Z 6.0: Decision Tree (excerpt; *coreferential* vs. all other categories)

5.3.2 Discussion

This section represents a discussion of the quantitative results presented in the previous section. First, aspects of the machine learning side of the task are discussed, including algorithms, experimental settings and feature sets; then, linguistic and cross-linguistic aspects are discussed.

5.3.2.1 Machine Learning Aspects

Regarding the different algorithms, decision trees seem particularly apt for the modelling of discourse-givenness. Obviously, there are no strong interactions between features, i.e. it is not a combination of a large number of features that tips the balance toward one or the other class. The finding that C4.5 generally performs a little better on this task than Ripper is consistent with Ng and Cardie (2002).

Another aspect that has proved important is the choice of the training data. Experiments with the same sets of features but different corpora and different methods for drawing subsets of these corpora were carried out. The use of different corpora, of course, yielded different ranges of results due to differences in the annotation schemes. This was also the case with the use of different subsets in MUC-7 (originally split into the sets Training, Dryrun and Formaleval). No differences were found, however, between results from training on data where documents were retained vs. data where instances were randomly distributed across sets (tested on OntoNotes and MUC-7 Formaleval)¹⁸. During the development process of a classifier, considering each of the experiments in isolation could have led to decisions for the further proceeding that contradict each other, for instance regarding the impact of certain feature groups. For instance, the best performing classifier is classifier 2 when trained on MUC-7 Train and testing on Dryrun, whereas it is classifier 3 when Formaleval is used instead for testing; when using MUC-7 Formaleval for training and Train+Dryrun for testing, it is classifier 4. In the present work, a comparison of the results under different settings has helped to minimize the effect of an overfitting to a certain training set.

The effect of the different feature groups can be summarised as follows: the baseline classifier produces relatively stable results across all different corpora, despite the differences in the size of the training set etc. ($60\% \pm 1\%$). As for the different groups of features, both the *local context* and the *newly introduced* features delivered a significant improvement on OntoNotes. On MUC-7, their effect is highly dependent on the choice of a subset for training. Here, it is probably advisable to take the results of the cross-validation experiment on all data (Train+Dryrun+Formaleval), assuming that the total amount of data describes the concept of coreference better than any of the subsets. From the results we observe that the *newly introduced* features contribute to an improvement in classifier performance. In ARRAU, it is also the *newly introduced* features that are crucial for improving the classification. Finally, this positive effect on classifier performance can also be shown on TüBa-D/Z. This holds for both groupings of categories (*coreferential/anaphoric/bound* vs. all other, as well as *coreferential* vs. all other). Thus, the usefulness of the *newly introduced* features from the areas of semantic similarity and

¹⁸The only exception is the relatively low performance of classifier 2 in the hold-out experiment on MUC-7 Formaleval compared to the 5-fold cross-validation experiment, see Table 5.11.

specificity for the classification of discourse-givenness of noun phrases has been shown for all corpora; the usefulness of *local context* features has been shown for coreference of specific entities in English (*specific* as defined in OntoNotes).

To get an idea how similar the models obtained by training are to the human annotator, consider the κ values in Tables 5.18 to 5.20.¹⁹

training:	Train	Train	5-fold
testing:	Dev	Test	cross-validation
1. baseline	0.55	0.54	0.55
2. no local context	0.70	0.69	0.71
3. no new	0.69	0.68	0.70
4. all	0.72	0.72	0.73

Table 5.18: OntoNotes 1.0: κ values for classifiers vs. original annotation

For OntoNotes, the values show consistency increasing with the use of additional features. They reach the level classified as ‘allowing for tentative conclusions’.

	5-fold cross validation	
	all sets	Formaleval
1. baseline	0.48	0.48
2. no local context	0.57	0.57
3. no new	0.53	0.57
4. all	0.56	0.57

Table 5.19: MUC-7: κ values for classifiers vs. original annotation

Consistency rises a little with additional features but is generally relatively low.

training:	Train	Train
testing:	Dev	Test
1. baseline	0.48	0.49
2. no local context	0.65	0.64
3. no new	0.60	0.61
4. all	0.64	0.62

Table 5.20: ARRAU 1.2: κ values for classifiers vs. original annotation

In ARRAU, consistency is quite low.

In TüBa-D/Z, for the first grouping of categories (*coreferential/anaphoric/bound* vs. all other categories), consistency reaches the level ‘allowing for tentative conclusions’ for classifiers trained on feature sets 2 to 4. The effect of additional features on consistency is relatively low. In the second grouping (*coreferential* vs. all other categories), additional features bring a massive improvement, but consistency is still low due to the difficult task.

¹⁹For the interpretation of κ values, see Section 3.2.3.2.

	coreferential/anaphoric/bound		coreferential	
training:	Train	Train	Train	Train
testing:	Dev	Test	Dev	Test
1. baseline	0.55	0.56	0.39	0.41
2. no local context	0.68	0.69	0.56	0.57
3. no new	0.68	0.69	0.54	0.56
4. all	0.68	0.70	0.56	0.58

Table 5.21: TüBa-D/Z 6.0: κ values for classifiers vs. original annotation

Inter-annotator agreement provides a measure of what can be expected of a classifier. As mentioned in Section 3.2.3.2, on MUC, agreement is reported to be 84% for precision and recall, which corresponds to 84% of f-measure, rising to 91% of f-measure if markable identification and coreference linking are carried out in separate steps (Hirschman et al., 1998). On OntoNotes, an average agreement between each annotator and the adjudicated results of 91.8% was yielded for the full coreference task (including antecedent identification). It can be assumed that a substantial proportion of the disagreement is due to disagreement on the antecedent, and that agreement on the discourse-givenness task is somewhat higher in the 90ies. On ARRAU, values around 0.6-0.7 (Krippendorff’s α) were obtained, and for TüBa-D/Z, agreement values of 83-85% f-measure for the full coreference task are reported.

A comparison of accuracy or κ values of the classification models to inter-annotator agreement gives the following picture: on MUC-7, the best classifier yields a weighted f-measure of 84.5%, which is comparable to human agreement (84%). On OntoNotes, accuracy reaches 93%. This is at least not lower than the performance of an average human performing the full coreference task (91.8%). A more precise assessment would be possible with detailed evaluation on each step and with f-measure values available. Regarding ARRAU, a comparison is not possible (α vs. κ). On TüBa-D/Z, weighted f-measure of the best classifiers (92-93%) is higher than inter-annotator agreement on the full coreference task (83%-85%). In summary, where a comparison is possible, the classifier’s performance reaches the level of human performance.

5.3.2.2 Linguistic Aspects

Across all corpora, it could be shown that an improvement in classification results could be obtained when using a combination of features modelling semantic similarity and ontological specificity. Local context features have been shown to have a positive effect in the case of OntoNotes, where discourse-givenness is restricted to noun phrases that are concrete and specific (the definition of *specific* is based on surface features), but includes bound expressions, as well as events given in the form of verbs. For the concept of discourse-givenness (and thus for coreference), this means that there are three components of the concept: explicit marking of specificity/vagueness, conceptual relatedness, and an object’s taking certain roles in a discourse. Tendencies to either of these factors can

be attributed to both the basic data and the definitions in the annotation schemes. News texts, for instance, often report on events with specific agents (usually persons or organizations). The schemes of OntoNotes and TüBa-D/Z concentrate on the marking of specific expressions, whereas in MUC-7 and ARRAU, generic (and in ARRAU also abstract) expressions are accepted for coreference.

Coreference in the strict sense, i.e. with specificity (defined in a superficial way) as a precondition, seems easier to model than coreference in the broad sense: while the baseline classifier performs similarly on all corpora (around 60% f-measure irrespectively of the size of the training set and of the class distributions), the additional features yield similar performance gains across corpora: on OntoNotes and TüBa-D/Z, both corpora defining specificity as a precondition for coreference, the gain is larger, despite the skewer class distribution.²⁰

From a practical point of view, I suggest to make the following compromises on the definition of coreference: (i) include binding, aggregations, and event anaphors into the annotation schemes and (ii) exclude presupposition anaphors, predication, apposition and function-value relations in order not to complicate the classification task (in other words, a union set of the expressions annotated in OntoNotes and TüBa-D/Z, without identity assertions).

As is observable from the error analyses, there is a number of rather infrequent phenomena, like idioms/collocations, expletives (in TüBa-D/Z also reflexives that are not inherent to the verb), coreference across boundaries of direct speech, aliases of names, metonymy, as well as shortened forms of referring expressions in headlines.

The former two, idioms/collocations and expletives/reflexives could be approached with distributional methods, i.e. using the main verb (and its other arguments, where existent), for deciding on the referentiality of an expression.

Remaining challenges include cataphors and entity-attribute relations. Expressions generally do not contain an explicit marking of whether they use the concepts they are composed from in a restrictive or an attributive way. This is what I consider the main challenge in the study of reference and coreference.

As is observable from the comparison between English and German, some of the features are not portable without adaption, e.g. *exact_previous_mention* (this issue can be solved with the help of lemmatization using e.g. Treetagger (Schmid, 1994)), and in particular the local context features. This is probably due to the differences in the realization of tense. A more sophisticated operationalization might help overcome this gap.

5.4 Comparison to Related Work

Table 4.4 in Section 4.4 gives an overview of results from related work; in this section, they are compared to my results.

²⁰Using the learning curve data, the training set size can be taken into account: compared to the MUC-7 classifiers from the 5-fold cross-validation experiment (using around 5,000 instances, yielding 65% f-measure), a classifier trained on OntoNotes yields over 70%, a classifier trained on TüBa-D/Z yields 66% using 5,000 instances. Compared to ARRAU (around 50,000 training instances), a classifier trained on OntoNotes yields over 74%, trained on TüBa-D/Z it yields also around 70%.

5.4.1 OntoNotes

There is one previous study on OntoNotes, Markert et al. (2012). This study presents a three-way categorisation (old, mediated, new) after having excluded non-referring expressions. They reuse OntoNotes’ original coreference annotation and add annotation according to Nissim’s (2004) annotation scheme. According to my calculations, OntoNotes originally contains a proportion of discourse-given instances of 15.78% of all NPs (including non-referring). Markert et al. (2012), after having excluded nonreferring expressions, and with adaptations to the annotation²¹, report proportions of 29.48% for old, 33.77% for mediated and 36.75% for new instances on a subset of 50 texts from the corpus (appr. 11,000 referring expressions). Their classifier reaches 85% of f-measure, but is not comparable.

5.4.2 MUC-7

Studies on MUC-7 include Ng and Cardie (2002) and Uryupina (2003, 2009). Ng and Cardie (2002) find a proportion of 26.8% of ‘old’ instances, which is roughly comparable to the proportion I calculated (25.08%). Uryupina (2003) finds 29% of ‘old’ instances in the Formaleval set; in her 2009 study, the ‘old’ instances make up approximately 34% of NPs in the Dryrun and in the Formaleval set. Differences are probably due to parsing and the mapping of coreference spans to noun phrases. Due to these differences, a direct comparison to Uryupina’s (2009) results is not quite unobjectionable.

Table 5.22²² shows the results already presented above. Unfortunately, Ng and Cardie only report their classifier’s accuracy, as their study is part of a larger work on coreference resolution.

It is noticeable that the tradeoff between accuracy and f measure (similar values for accuracy, different values for f measure of the smaller class) is different between Uryupina’s (2009) work and this work. I attribute this to the differences in the original class distributions (the proportion of discourse-given instances in the set Dryrun+Formaleval is 34% according to Uryupina (2009), 26.13% in this work). The weighted average values of f measure, however, are very similar (Uryupina 84.74%, this work 84.00%).

In both settings, the classifiers are trained on a relatively small number of instances (around 5,000 in the first setting, around 3,000 in the second). In the first setting (where values are averaged), the standard deviation in this work is 15% of f measure of the smaller class for classifiers 1 and 3, 13% for classifier 2 and 11% for classifier 4. Given this deviation, conclusions drawn based on results of such small data sets should be treated with caution.

5.4.3 Related Work in General

Unfortunately, a comparison to related work is difficult: the tasks are very similar, but not identical. Most of the work on the classification of information status assumes that non-referring expressions have been excluded beforehand.

²¹According to Nissim’s annotation scheme, situationally given expressions are also annotated as *old*.

²²Numbers marked with asterisks ‘*’ represent values calculated a posteriori.

	acc	P	R	F	class
hold out one document at a time (Formaleval)					
Uryupina (2003)	ø81.1%*	ø65.8%*	ø73.4%*	ø69.4%*	-discourse-new
		ø88.5%	ø84.3%	ø86.3%	+discourse-new
this work					
1. baseline	83.04%	77.8%	47.1%	58.0%	+discourse-given
		83.5%	95.8%	89.2%	-discourse-given
2. no local context	83.92%	74.6%	55.7%	58.1%	+discourse-given
		85.6%	93.9%	89.5%	-discourse-given
3. no new	83.59%	72.7%	56.4%	62.7%	+discourse-given
		85.9%	93.1%	89.3%	-discourse-given
4. all	84.68%	76.0%	59.1%	66.0%	+discourse-given
		86.5%	93.9%	89.9%	-discourse-given
trained on Dryrun, tested on Formaleval					
Ng and Cardie (2002)	84.0%				+discourse-given
					-discourse-given
Uryupina (2009)	84.4%*	71.6%*	88.7%*	79.2%*	-discourse-new
		93.5%	82.3%	87.6%	+discourse-new
this work					
1. baseline	82.97%	79.3%	47.9%	59.7%	+discourse-given
		85.9%	92.7%	89.2%	-discourse-given
2. no local context ^{1***}	84.99%	78.8%	58.8%	67.4%	+discourse-given
		86.5%	94.3%	90.3%	-discourse-given
3. no new ^{2**}	83.90%	82.3%	49.5%	61.8%	+discourse-given
		84.2%	96.2%	89.8%	-discourse-given
4. all ^{1***}	84.96%	81.2%	55.8%	66.2%	+discourse-given
		85.8%	95.4%	90.3%	-discourse-given

Table 5.22: MUC-7: Comparison of Classification Results

For a rough comparison of these works to my experiments, Markert et al.’s results of an information status classification model (trained on 9 of 10 folds of around 10,000 NPs from OntoNotes) can be juxtaposed to my results of the discourse-givenness classification model trained on around 31,000 NPs from ARRAU (the basic data of which largely overlaps with OntoNotes). Whereas Markert et al.’s model reaches around 85% f-measure for the class *old* and 79% accuracy (see Section 4.4), my model reaches around 79% f-measure for the class *discourse-given* and 85% accuracy (see Section 5.3.1.3 for more details). A more meaningful comparison remains to be done, focussing on the overlapping data.

The main differences to work on the classification of discourse-givenness are differences in the experimental setup: Uryupina’s and Ng and Cardie’s work differs in preprocessing like parsing and mapping, and resulting from that, in the training on sets of different sizes. Cahill and Riestler’s work differs in that they make use of (automatically detected or manually annotated gold standard) coreference information.

In contrast to previous studies, in the present work, feature sets have been tested on several corpora, and their impact can be set into relation to the concepts annotated.

Chapter 6

Conclusions and Outlook

In the first part of this work, the concept of *discourse-givenness* was examined from a theoretical perspective. In the second part, several resources, each of them instancing a different definition, were used to build computational models of *discourse-givenness*. Applying the same methods to all resources and analyzing the respective results allows to relate the methods to the annotation definitions of the respective resources. From this, conclusions can be drawn that might be fed back into the advancement of theoretical definitions. These conclusions are presented in the following.

6.1 Conclusions

From the analysis of the theoretical work in Chapter 2, it becomes obvious that the concepts of *discourse-givenness* and *coreference* are not yet at a stage of being exhaustively defined or even algorithmized,¹ mainly due to open issues in the definition of *reference* (see Section 2.2). Yet, the analysis of resources annotated with these concepts (Chapter 3) shows a general consensus that a subset with a clearer-cut definition exists, namely coreference presupposing specific reference.² An overview of related work is given in Chapter 4, showing some obstacles in the way of comparability of approaches. In Chapter 5, I introduced features representing (i) how specific (or abstract, respectively) an expression is, (ii) how similar single words contained in the expression are to the words in the previous context, and (iii) hints the (local) context provides on the expression's discourse-givenness. Features representing properties (i) and (ii) proved useful on all corpora. Features representing property (iii) proved useful for the following settings: distinguishing specific coreferent expressions from other NPs (see Section 5.3.1.1 for more details) in English. In general, the concept of specific coreference was learnt better by the classifiers than the concept of coreference including generic and more abstract reference. The distinction of bound vs. coreferential (experiments on TüBa-D/Z) remains a challenge.

For the theory, this means that *reference* (in particular, non-specific reference) needs to

¹Attempts, including Heim and Kratzer (1998) (from p. 261) and Buring (2005), mainly focus on pronouns.

²Focussing on a clear-cut subset is practiced for instance in the ACE corpus (see Section 3.1.1), which only contains coreference for certain predefined entities. This corpus, however, includes predication (see Section 2.3.4) for reasons of the application it is geared to.

be defined more precisely. In particular, the open issues are: which kinds of expressions may refer? How do we cope with possible worlds? How do we account for the context implicitly narrowing the referent set (see Example (125) repeated from (33); the NP *the nation* narrows *Highway officials* to *U.S. highway officials* here; it could even be argued that *Everyone* is narrowed to *everyone in (or from) the U.S.*)? How do we determine the scope of generalizations? (See Section 2.2 for a discussion of these issues.)

- (125) Everyone agrees that most of **the nation's** old bridges need to be repaired or replaced. But there's disagreement over how to do it. *Highway officials* insist the ornamental railings on older bridges aren't strong enough to prevent vehicles from crashing through.

In *coreference*, the open issues are: how do we analyze discourses referring to different aspects of the same object? How do we analyze discourses referring to different (temporal) stages of the same object? Does coreference hold between generic NPs? (See Section 2.3 for a discussion of these issues.)

Linguistic diagnostics for *reference*, *specificity* and *coreference* are still a desideratum, as can be seen from the operationalizations in the annotation guidelines (in particular, the specific/generic distinction, to the effect that binding and coreference are practically conflated), see Chapter 3. Based on the finding that the local context helps determining specific coreference in English, diagnostics could be constructed using the local context (e.g. tense of the main verb in comparison to previous sentences; continuity of meaning after change of tense etc.).

Finally, it seems that in coreference resolution, a nominal referring expression's informational content has been considered sufficient for resolving what it refers to. This might have to be reconsidered.

6.2 Outlook

In this section, lines of further research will be sketched. In particular, these include possible improvements to the current system for discourse-givenness classification, general suggestions for classification approaches in the field of discourse-givenness and information status, as well as methods from neighboring disciplines that might help to gain new insights on the concept of discourse-givenness, its cognitive processing and acquisition.

The existing classification system could be improved with more sophisticated operationalizations for the verb tense features (currently, suffixes of verbs are used). A rule based preprocessing determining the tense of the clause's main verb is likely to yield more precise results, though at higher cost (manual effort for the declaration of rules and computation time). Another improvement could be made with the use of selected verb lemmas that tend to subcategorize expletives (or reflexives), or with recurrent idioms and phrases (e.g. *lead the way*, *have the option*, *be the subject*).

For different modalities (written vs. spoken data, monolog vs. dialog), separate models should be trained.

For distinguishing bound expressions from coreferential expressions (see Section 5.3.1.4, Table 5.17 for a preliminary study), the additional use of an expression's neighbor expressions and their features seems to be indicated.

Regarding the classification of discourse-givenness and information status in general, it seems natural to test a cascade architecture. In the first step, non-referring expressions are excluded. For this step, ARRAU could be used as training material, as it contains referentiality annotation. In the second step, new (referring) expressions are excluded. The resulting expressions can be passed to a system for coreference resolution. Evidence that a breaking down of tasks might be successful is provided by Hirschman et al. (1998), who observed an increase in human inter-annotator agreement after breaking down the task into a first step of markable identification and a second step of coreference linking (cf. also (Goecke et al., 2007)).

Aiming at a better understanding of discourse-givenness in general, experiments in the neighboring fields of psycholinguistics and language acquisition could help to gain insights to the following questions:

There seem to be two concurrent strategies for resolving the reference of locational and temporal expressions: deictic (corresponding to instant accomodation) and coreferent. Does deixis/accomodation overrule coreference, i.e. are temporal and locational expressions perceived primarily as deictic or as coreferent? Do temporal and locational referents have a shorter span of activation than other referents? The fact that some of the annotation errors found during the error analyses were locational or temporal expressions (*last year, next year, the state*, etc.), as well as the fact that in many studies on inter-annotator agreement Section 3.2.3.2), such expressions were explicitly excluded, suggests that these matters need further investigation.

Studies on the cognitive processing and the acquisition of coreference resolution (both in children and in second language learners) could provide clues for an algorithmization (or at least for an ordered set of heuristics) of coreference resolution. In particular, what cues do humans use to determine whether an expression is specific or generic? (Eye-tracking could be used to reconstruct which parts of sentences are used during the decision process.) Do humans perceive repeated reference to the same kind as coreference? What are the exact conditions for entity-attribute relations (e.g. can an entity be taken up using one of its attributes several sentences after its mention, is it necessary that the lexical content of the expressions be related)? From which age can a human differentiate expressions that are bound from expressions that are coreferential? When does a human acquire the ability to resolve entity-attribute relations (as opposed to pronomial coreference)? Are there differences between languages with respect to the syntactic conditions under which these relations can occur?

With these issues resolved, advancements in the classification of discourse-givenness and information status, as well as in coreference resolution can be expected.

Chapter 7

Summaries

This chapter provides a short summary of each chapter. An index for cross-reading is provided at the end of the document.

7.1 Summary in English

Introduction

A central component in text understanding is coreference resolution, i.e. finding those expressions in a text that relate to the same object in the world. A part of coreference resolution is the distinction of discourse-given noun phrases. A noun phrase is discourse-given if it refers to something mentioned in the previous context. In the work at hand, concepts like reference, coreference, discourse-givenness and information status are defined based on Discourse Representation Theory (DRT, Kamp and Reyle (1993)). The goal of this work is to build a theoretically well-founded computational model of discourse-givenness for English and German using Machine Learning methods. Possible applications include automatic coreference resolution, speech synthesis, and a deeper linguistic discourse analysis.

Theoretical Background

Reference, Coreference, Discourse-givenness and Information Status are concepts that have been widely discussed in the literature. However, there is disagreement on the definition criteria as well as on the terminology, in particular regarding the concept of *reference*. Formal definitions are available only for some of these concepts.

Criteria for the definition of reference once postulated, had to be given up (identifiability of the referent by the speaker, for instance, or the existential presupposition of referring expressions, see von Heusinger (2002)). Reference in its broader sense includes reference to kinds or abstract concepts, while in its strict sense, this is excluded. Expressions referring in the strict sense are also termed specific or definite. These terms however, are used in their strict sense to refer to noun phrases with a definite determiner, proper names and non-bound pronouns. In this work, the definitions discussed in Bach (1987) are taken up and formalised. Open questions in the identification and distinction of referring expressions are pointed out with the help of examples, for instance, a delimitation of

potentially referring expressions by means of their syntactic category (Chomsky, 1965) or the issue of possible worlds (to the consequence that the recipient cannot be sure about the world of evaluation).

A simplified version of Bach’s definition of reference – reference as a function – is used in definitions of coreference as a relation between expressions with identical referents (Kamp and Reyle, 1993; Kibble and van Deemter, 2000). Discourse-givenness is then defined as a constituent’s having a coreferent expression in the context preceding it (Riester, 2009). These definitions in conjunction allow for a representation of the content of a discourse. This content, however, discloses itself only after the interpretation of the discourse. The interpretation is partly dependent on world knowledge. An algorithm for the interpretation process, for identifying referring expressions and resolving coreference, remains a desideratum.

Consulting DRT as a formalisation framework, however, serves the purpose of distinguishing coreference from other phenomena. In particular, it is a necessary precondition for discourse-givenness that a variable has been introduced into the main DRS. Only there, it is available for being “docked on” by another referent’s variable using the identity relation. If, for postulating identity, such a variable has to be generated (for instance, by *summation*), the relation is not a coreference relation.

Some of the relations similar to coreference are subsumed under categories that form part of the concept of information status. Information status includes the categories *discourse-given*, *inferable* (also called *accessible* or *mediated*, meaning situatively or indirectly given, for instance via *bridging*, *summation* etc.) and *new* (any other expression, except non-referring expressions).

As a result, this chapter contains definitions of *reference*, *coreference* and *discourse-givenness* that are more precise than in previous approaches.

Existing Resources

Machine Learning models need large amounts of training instances. This training data can be found in existing corpora annotated with coreference and/or information status. After having surveyed the theoretical side, I investigated how the concepts have been implemented in practice. The following corpora are reviewed in detail: for English, the Message Understanding Conference 7 corpus (MUC-7), OntoNotes, Nissim’s Nissim et al. (2004) annotations of the Switchboard corpus and ARRAU; for German, the Potsdam Commentary Corpus (PCC), the Tübingen Treebank (TüBa-D/Z) and the Discourse Information Radio News Database for Linguistic analysis (DIRNDL). The analysis makes evident that annotation schemes pursue different paths with respect to phenomena like coreference between generic expressions, as well as event and proposition anaphors. Also, some of the schemes make simplifications for the sake of consistency of annotation, for instance, dropping the distinction between coreference and binding.

Related Work

There is about a dozen of related studies, many of them based on MUC or its successor ACE. Most studies are concerned with the finding of useful features. New features have been sought mainly in the areas of syntactic form and structure, as well as semantic

relatedness to preceding instances or agreement with respect to the semantic class. As is the case with many tasks, discourse givenness classification relies in large parts on rather simple features, for instance, ‘head_previous_mention’, a feature counting the number of times the NP’s head lemma has been used in its previous context. The solution is thus to be expected rather in a number of smaller improvements than in one measure capturing all instances.

Classification algorithms repeatedly used include decision trees and support vector machines, among others. Recently, sequential classification methods have been used, which also take an instance’s neighbors into account. Direct comparisons are difficult due to differences in the annotation schemes and sizes of training sets. Even works on the same corpus are hard to compare due to differences in preprocessing (MUC, for instance, does not include noun phrase boundaries). In some studies (Nissim (2006) and followers), non-referring expressions are excluded in advance, which simplifies, or at least changes the classification task to a certain extent.

Discourse-Givenness Classification: New Experiments and Results

In the present work, the English corpora MUC-7, OntoNotes 1.0 and ARRAU and the German corpus TüBa-D/Z are used to train classifiers for discourse-givenness. The class, +discourse-given or -discourse-given, is defined by the original coreference annotation of each of the corpora. For the modelling, features from the literature are adapted on the one hand. On the other hand, new features are added from the following areas: semantic similarity between expressions, ontological specificity¹ of the concepts used in the respective noun phrase, and the local context of the respective noun phrase. The resulting classifier is compared to a baseline classifier, which uses features comparable to Nissim (2006). Across all corpora, the additional features measuring specificity and similarity to previously mentioned entities have significantly contributed to the improvement of classification results, in particular a rise in recall (i.e. of the discourse-given entities, a higher proportion can be found with the new features). Features describing the local context of a noun phrase proved helpful in the distinction of specific discourse-given expressions in OntoNotes. In general, the classifiers reach the level of performance of human annotators where a comparison is possible.

Conclusions and Outlook

Given that better results were achieved for specific coreference (OntoNotes and TüBa-D/Z) than for coreference in a broader sense (MUC-7 and ARRAU), it can be concluded that clues for coreference between (superficially) non-specific expressions need a more precise definition. A couple of improvements to the system have been proposed, including a refinement of local context features. Further, a cascade model has been suggested, where the non-referring expressions are filtered out in the first pass, discourse-new expressions are filtered out in the second pass, and the remaining expressions would be handed over

¹Here, the term *specificity* is used in the sense of *strength of distinction* (*specific* as opposed to *common*) and is a property of concepts. It is not to be confused with *specificity* in the sense of *semantic definiteness*, a property of noun phrases used above (*specific* as opposed to *vague* or *anonymous*).

to a coreference resolution component. Finally, an outline is given how research on *discourse-giverness* can make use of methods from neighboring disciplines to answer the questions identified in this work.

7.2 Summary in German (*Zusammenfassung in deutscher Sprache*)

Es folgt eine kapitelweise Zusammenfassung der vorliegenden Arbeit. Ein Index am Ende des Dokuments soll das gezielte Nachlesen ausgewählter Passagen erleichtern.

Einführung

Einen zentralen Bestandteil des Textverstehens bildet die Koreferenzresolution, also das Identifizieren derjenigen sprachlichen Ausdrücke innerhalb eines Textes, die sich auf dieselben Objekte in der realen Welt beziehen. Eine Komponente davon wiederum bildet das Auffinden diskursgebener Nominalphrasen, d.h. derjenigen Nominalphrasen, die sich zurückbeziehen auf bereits Erwähntes. In der vorliegenden Arbeit werden die jeweils aufeinander aufbauenden Konzepte *Referenz*, *Koreferenz*, *Diskursgegebenheit* und *Informationsstatus* definiert, und zwar unter Rückgriff auf Kamp und Reyle's (1993) Diskursrepräsentationstheorie (DRT). Das Ziel der Arbeit ist eine theoretisch fundierte Modellierung von Diskursgegebenheit mit Hilfe von Verfahren des Maschinellen Lernens. Motiviert ist die Modellierung durch mögliche Anwendungen in Systemen zur automatischen Koreferenzresolution, der Sprachsynthese und der linguistischen Diskursanalyse.

Theoretischer Hintergrund

Die Konzepte *Referenz*, *Koreferenz*, *Diskursgegebenheit* und *Informationsstatus* werden in einer Vielzahl von Veröffentlichungen diskutiert. Allerdings herrscht Uneinigkeit über Definitionskriterien und Terminologie, v.a. bezüglich des grundlegenden Konzeptes *Referenz*. Formale Definitionen existieren nur zum Teil.

Definitionskriterien für Referenz, die zunächst postuliert wurden, wurden wieder aufgegeben, z.B. die Identifizierbarkeit des Referenten durch den Sprecher und die Existenzpräsupposition von referierenden Ausdrücken (von Heusinger, 2002). Während der Term Referenz im weiteren Sinne auch Ausdrücke einschließt, die Arten von Objekten (*kinds*) oder abstrakte Konzepte beschreiben, sind diese Ausdrücke bei der strikten Definition von Referenz ausgeschlossen. Im strikten Sinne referierende Ausdrücke werden gelegentlich auch als spezifisch oder definit bezeichnet; allerdings werden die letzteren beiden Begriffe wiederum im strikten Sinne verwendet, um Nominalphrasen mit definitem Artikel, Eigennamen und Pronomina mit konkretem Bezug zu bezeichnen. In der vorliegenden Arbeit werden die in Bach (1987) diskutierten Definitionen aufgegriffen und formalisiert. Anhand von Beispielen werden einige offene Fragen in der Identifikation und Abgrenzung von referierenden Ausdrücken aufgezeigt, z.B. eine Eingrenzung von Ausdrücken, die referieren können, hinsichtlich der syntaktischen Kategorie (Chomsky,

1965), sowie die Problematik möglicher Welten (mit der Folge der Unklarheit über das Auswertungsuniversum für den Rezipienten).

Aufbauend auf einer simplifizierten Version der Bachschen Definition von Referenz – der Definition von Referenz als Funktion – wird Koreferenz definiert als eine Relation zwischen Ausdrücken mit identischen Referenten (Kamp and Reyle, 1993; Kibble and van Deemter, 2000). Die Diskursgegebenheit einer Konstituente ist definiert über das Vorhandensein eines koreferenten Ausdrucks im vorangehenden Kontext (Riester, 2009). Diese Definitionen erlauben zwar die Repräsentation des Gehalts eines Diskurses, der sich aus dessen Interpretation ergibt. Ein Algorithmus für den Interpretationsschritt zur Erkennung und Auflösung von Koreferenz – der z.T. Weltwissen erfordert – bleibt jedoch ein Desiderat.

Nichtsdestotrotz ist das Hinzuziehen der DRT insofern sinnvoll, als diese Formalisierung eine eindeutige Abgrenzung von Koreferenz gegen andere Phänomene ermöglicht: Diskursgegebenheit setzt voraus, dass eine Variable in der Haupt-DRS bereits eingeführt ist, an die eine neue Referentenvariable mit der Identitätsrelation “andocken” kann. Befindet sich diese Variable nicht in der Haupt-DRS oder muss, um Identität herzustellen, erst eine Variable gebildet werden (z.B. durch *Summation*, d.h. Zusammenfassen mehrerer Einzelreferenten), liegt keine Koreferenz im strikten Sinne vor.

Einige der Koreferenz ähnliche Relationen werden schließlich unter der Kategorie Informationsstatus erfaßt. Dieses Konzept beinhaltet die Unterkategorien *diskursgegeben*, *erschließbar* (d.h. situativ oder indirekt gegeben, z.B. durch *Bridging*, *Summation* etc.) und *neu* (alle anderen, ausschließlich der nichtreferentiellen).

Im Ergebnis stellt diese Kapitel eine Präzisierung bisheriger Definitionen von *Referenz*, *Koreferenz* und damit auch *Diskursgegebenheit* dar.

Existierende Ressourcen

Zur Modellierung mittels maschineller Lernverfahren werden größere Mengen von Trainingsdaten benötigt. Hier bietet es sich an, auf existierende Ressourcen zurückzugreifen.

In diesem Teil der Arbeit wird analysiert, wie die theoretischen Konzepte bei der Korpusannotation in die Praxis umgesetzt wurden. Insbesondere werden die folgenden Korpora detailliert besprochen: die Korpora der Message Understanding Conference 7 (MUC-7), OntoNotes, Nissim’s (2004) Annotationen des Switchboard Korpus und ARRAU fürs Englische, sowie das Potsdamer Kommentarkorpus (PCC), die Tübinger Baubank Deutsch/Zeitungssprache (TüBa-D/Z) und die Discourse Information Radio News Database for Linguistic analysis (DIRNDL, basierend auf dem Stuttgarter Radionachrichtenkorpus) fürs Deutsche. Bei der Analyse zeigt sich zum einen, dass sehr unterschiedliche Entscheidungen getroffen wurden zu Phänomenen wie z.B. Koreferenz zwischen generischen Ausdrücken, sowie Ereignis- und Propositionsanaphern. Das erschwert Vergleiche zwischen Modellen, die auf diesen unterschiedlichen Daten basieren. Zum anderen zeigt sich, dass zugunsten der Annotationskonsistenz zum Teil Vereinfachungen vorgenommen wurden, z.B. werden nach manchen Annotationsschemata Fälle von *Binding* als Fälle von Koreferenz behandelt.

Stand der Forschung

Ein großer Teil der Arbeiten zur Klassifikation von Diskursgegebenheit basiert auf dem MUC-Korpus bzw. seinem Nachfolge-Korpus ACE. Die meisten Studien beschäftigen sich mit dem Auffinden geeigneter Merkmale – diese werden vor allem gesucht im Bereich der syntaktischen Form und Struktur, sowie der semantischen Verwandtschaft oder Übereinstimmung hinsichtlich der semantischen Klasse. Wie bei vielen Problemen zeigt sich auch bei Diskursgegebenheit, dass ein großer Teil der Fälle mit relativ simplen Mitteln abgedeckt werden kann, z.B. mit dem Merkmal ‘head_previous_mention’, das festhält, wie oft das Lemma des Kopfes einer NP im linken Kontext bereits verwendet wurde. Es ist zu erwarten, dass hier mehrere kleine Ergänzungen gemacht werden müssen, um eine Verbesserung zu erzielen.

Häufig verwendete Klassifikationsalgorithmen sind unter anderem Entscheidungsbäume und Support Vector Machines. Neuerdings werden auch sequentielle Methoden eingesetzt; diese berücksichtigen bei der Klassifikation einer Instanz auch die benachbarten Instanzen.

Ein Vergleich der bestehenden Verfahren ist äußerst schwierig: zwischen verschiedenen Korpora unterscheiden sich die Annotationsschemata stark; auch die Anzahl der Trainingsinstanzen ist unterschiedlich. Bei Verfahren, die auf denselben Daten arbeiten, ergeben sich z.T. durch die Vorverarbeitung Unterschiede, z.B. im Falle von MUC, das keine Syntaxannotation enthält und damit keine Nominalphrasengrenzen. In einigen Arbeiten (Nissim (2006) und daran anschließende Arbeiten) werden nichtreferentielle Ausdrücke vorab ausgeschlossen, was zu einer leichten Veränderung der Klassifikationssaufgabe führt.

Neue Experimente und Ergebnisse zur Klassifikation von Diskursgegebenheit

In der vorliegenden Arbeit werden Klassifikatoren für Diskursgegebenheit auf den Korpora MUC-7, OntoNotes 1.0 und ARRAU (fürs Englische) und TüBa-D/Z 6.0 (fürs Deutsche) trainiert, wobei die jeweilige Original-Koreferenzannotation die Zielklasse jeder Nominalphrase (diskursgegeben oder nicht diskursgegeben) definiert. Zur Modellbildung werden einerseits Merkmale, die in der Literatur beschrieben sind, übernommen. Diese werden ergänzt durch neue Merkmale, die sich in folgende drei Bereiche gliedern lassen: Ähnlichkeit zu Ausdrücken im Kontext, ontologische Spezifität² der in der Nominalphrase verwendeten Konzepte, sowie der lokale Kontext der Nominalphrase. Zum Vergleich wird ein Baseline-Klassifikator herangezogen, der auf Merkmalen basiert, die mit Nissim (2006) vergleichbar sind. Auf allen Korpora konnten durch das Hinzufügen neuer Merkmale aus den Bereichen Spezifität und Ähnlichkeit zum vorangehenden Kontext Verbesserungen im Vergleich zur Baseline erzielt werden, v.a. im Hinblick auf den Recall diskursgegebener Entitäten (d.h. von den diskursgegebenen Entitäten wurde ein höherer Anteil identifiziert als bisher). Merkmale, die den lokalen Kontext einer Nom-

²Der Begriff *Spezifität* wird hier im Sinne von *Genauigkeit der Charakterisierung* verwendet (*spezifisch* in Abgrenzung zu *allgemein*). Dabei handelt es sich um eine Eigenschaft von Konzepten und ist nicht zu verwechseln mit *Spezifität* im oben verwendeten Sinn von *semantischer Definitheit* als Eigenschaft von Nominalphrasen (*spezifisch* in Abgrenzung zu *vage* oder *anonym*).

inalphrase näher charakterisieren, erzielten Verbesserungen bei der Klassifikation von spezifischen diskursgegebenen Ausdrücken in OntoNotes. Insgesamt erreichen die Klassifikatoren Ergebnisse, die sich mit denen menschlicher Annotatoren messen können.

Fazit und Ausblick

Aufgrund der Tatsache, dass bei der Klassifikation spezifischer Koreferenz (auf der Basis von OntoNotes und TüBa-D/Z) bessere Resultate erzielt wurden als bei der Klassifikation von Koreferenz im weiteren Sinne (auf der Basis von MUC-7 und ARRAU), komme ich zu dem Ergebnis, dass Kriterien für Koreferenz zwischen nicht-spezifischen Ausdrücken noch feiner ausgearbeitet werden müssen. Für das bestehende Klassifikationssystem wurden Verbesserungsvorschläge gemacht, z.B. eine differenziertere Version für einige der Merkmale, die den lokalen Kontext beschreiben. Zudem wurde ein Kaskadenmodell skizziert, bei dem im ersten Schritt die nichtreferentiellen Ausdrücke, im zweiten die diskursneuen herausgefiltert werden, sodass schließlich die verbliebenen, diskursgegebenen Ausdrücke an eine Koreferenzresolutionskomponente übergeben werden können. Zum Schluß des Ausblicks wird aufgezeigt, wie die (computer-)linguistische Forschung im Bereich Diskursgegebenheit von den Methoden in angrenzenden Disziplinen wie der Psycholinguistik und der Spracherwerbsforschung profitieren kann, um zu Fortschritten hinsichtlich der aufgezeigten offenen Fragen zu gelangen.

References

- Albrecht, Irene, Schröder, Marc, Haber, Jörg, and Seidel, Hans-Peter. 2005. Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality*, 8:201–212, August.
- Allan, Keith. 2012. Referring to 'what counts as the referent': a view from linguistics. In Capone, Alessandro, Piparo, Franco Lo, and Carapezza, Marco, editors, *Perspectives on Pragmatics and Philosophy*. Springer Verlag, Milan.
- Artstein, Ron and Poesio, Massimo. 2005. $\text{Kappa}^3 = \text{Alpha}$ (or Beta). Technical report, Department of Computer Science, University of Essex, UK.
- Asher, Nicholas and Lascarides, Alex. 1998. The Semantics and Pragmatics of Presupposition. *Journal of Semantics*, 15:239–299.
- Asher, Nicholas and Lascarides, Alex. 2003. *Logics of Conversation*. Cambridge University Press.
- Asher, Nicholas. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Boston MA.
- Authorless. 2007. Co-reference Guidelines for English OntoNotes. Version 6.0.
- Bach, Kent. 1987. *Thought and Reference*. Clarendon Press, Oxford.
- Baumann, Stefan and Riester, Arndt. 2012. Referential and Lexical Givenness: Semantic, Prosodic and Cognitive Aspects. *Prosody and Meaning (Trends in Linguistics)*, 25 of *Interface Explorations*:119–162.
- Behnke, Heinrich, Bachmann, Friedrich, Fladt, Kuno, and Süß, Wilhelm, editors. 1958. *Grundzüge der Mathematik*. Vandenhoeck & Ruprecht, Göttingen.
- Bergsma, Shane, Lin, Dekang, and Goebel, Randy. 2008. Distributional Identification of Non-Referential Pronouns. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL2008:HLLT)*, pages 10–18, Columbus, Ohio.
- Bonafonte, Antonio, Aguilar, Lourdes, Esquerra, Ignasi, Oller, Sergio, and Moreno, Asunción. 2009. Recent work on the FESTCAT database for speech synthesis. In *Iberian SLTech 2009*, pages 131–132.
- Brants, Sabine, Dipper, Stefanie, Hansen, Silvia, Lezius, Wolfgang, and Smith, George. 2002. The TIGER Treebank. In Hinrichs, Erhard and Simov, Kiril, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, pages 24–41, Sozopol, Bulgaria.

- Brants, Thorsten. 2000. Tnt - a statistical part-of-speech tagger. In *6th Applied Natural Language Processing (ANLP '00), April 29 - May 4*, pages 224–231, Seattle, USA. Association for Computational Linguistics.
- Brown, Gilian. 1983. Prosodic structure and the given/new distinction. In Cutler, A. and Ladd, D.R., editors, *Prosody: Models and measurements*, pages 67–77. Springer Verlag, Berlin.
- Büring, Daniel. 2005. *Binding Theory*. Cambridge Textbooks in Linguistics. Cambridge University Press, Cambridge.
- Büring, Daniel. 2007. Binding. To appear in Patrick Hogan, editor, *The Cambridge Encyclopedia of the Language Sciences*.
- Bußmann, Hadumod. 2000. *Lexikon der Sprachwissenschaft*. Kröner, Stuttgart.
- Byron, Donna and Gegg-Harrison, Whitney. 2004. Eliminating non-referring noun phrases from coreference resolution. In *Proceedings of DAARC*, pages 21–26.
- Byron, Donna K. 2002. Resolving Pronominal Reference to Abstract Entities. In *Proceedings of the 2002 annual meeting of the Association for Computational Linguistics (ACL2002)*, pages 80–87, Philadelphia, PA, USA.
- Cahill, Aoife and Riester, Arndt. 2012. Automatically Acquiring Fine-Grained Information Status Distinctions in German. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 232–236, Seoul, South Korea, July. Association for Computational Linguistics.
- Calhoun, Sasha, Nissim, Malvina, Steedman, Mark, and Brenier, Jason M. 2005. A Framework for Annotating Information Structure in Discourse. In *Proceedings of the ACL 2005 Workshop on Frontiers in Corpus Annotation II: Pie in the Sky*, pages 45–52, Ann Arbor, Michigan. Association for Computational Linguistics.
- Carletta, Jean. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.
- Carlson, Gregory N. 1977a. A Unified Analysis of the English Bare Plural. *Linguistics and Philosophy*, 1:413–456.
- Carlson, Gregory N. 1977b. *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts, Amherst.
- Carlson, Gregory N. 1989. The semantic composition of english generic sentences. In Chierchia, Gennaro, Partee, Barbara Hall, and Turner, Raymond, editors, *Properties, Types and Meaning: Semantic Issues*, pages 167–191. Kluwer, Dordrecht.
- Cassell, Justine, Vilhjálmsón, Hannes Högni, and Bickmore, Timothy. 2001. BEAT: the Behavior Expression Animation Toolkit. In *Proceedings of SIGGRAPH*, pages 477–486, New York. ACM.
- Chafe, Wallace L. 1970. *Meaning and the Structure of Language*. The University of Chicago Press, Chicago.
- Chafe, Wallace L. 1976. Givenness, contrastiveness, definiteness, subjects, topics and point of view. In Li, Charles N., editor, *Subject and Topic*, pages 27–55. Academic

- Press, New York.
- Chafe, Wallace L. 1980. The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. *Advances in Discourse Processes*, 3:323. (editor).
- Charniak, Eugene. 2000. A maximum-entropy-inspired parser. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, pages 132–139.
- Chiarcos, Christian and Krasavina, Olga. 2005. Annotation Guidelines PoCoS Potsdam Coreference Scheme: Core Scheme. ms. (Draft version 0.912).
- Chiarcos, Christian and Ritz, Julia. 2010. Qualitative and Quantitative Error Analysis in Context. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, September.
- Chinchor, Nancy and Sundheim, Beth. 2003. Message Understanding Conference (MUC) 6.
- Chinchor, Nancy. 1998. MUC-7 named entity task definition (version 3.5). In *Proceedings of the Seventh Message Understanding Conference*, Fairfax, Virginia. Science Applications International Corporation.
- Chinchor, Nancy. 2001. Message Understanding Conference (MUC) 7.
- Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, Mass.
- Clark, Herbert H. 1975. Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, TINLAP '75, pages 169–174, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Clément, Danièle. 2000. Linguistisches Grundwissen. In *Einführung in die germanistische Linguistik*. Westdeutscher Verlag, Wiesbaden, 2 edition.
- Cohen, Ariel and Erteschik-Shir, Nomi. 2002. Topic, Focus, and the Interpretation of Bare Plurals. *Natural Language Semantics*, 10(2):125–165.
- Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.
- Cohen, William W. 1995. Fast Effective Rule Induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.
- Cohen, Ariel. 2001. On the generic use of indefinite singulars. *Journal of Semantics*, 18(3):183–209.
- Curran, James R. and Clark, Stephen. 2003. Language independent NER using a maximum entropy tagger. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, CONLL '03, pages 164–167, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Dahl, Östen. 1975. On Generics. *Formal Semantics of Natural Language*, pages 99–111.
- Denis, Pascal and Baldrige, Jason. 2007. Joint Determination of Anaphoricity and Coreference Resolution using Integer Programming. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 236–243, Rochester,

- New York, April. Association for Computational Linguistics.
- Dipper, Stefanie and Zinsmeister, Heike. 2010. Towards a standard for annotating abstract anaphora. In *Proceedings of the LREC workshop on Language Resource and Language Technology Standards*, pages 54–59, Valletta, Malta.
- Doddington, George R., Mitchell, Alexis, Przybocki, Mark, Ramshaw, Lance, Strassell, Stephanie, and Weischedel, Ralph. 2000. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of Conference on Language Resources and Evaluation (LREC 2004)*.
- Donnellan, Keith S. 1966. Reference and Definite Descriptions. *The Philosophical Review*, 75(3):281–304.
- Eckart, Kerstin, Riestler, Arndt, and Schweitzer, Katrin. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In et al., Christian Chiarcos, editor, *Linked Data in Linguistics*. Springer, Berlin.
- Elsner, Micha and Charniak, Eugene. 2007. A Generative Discourse-New Model for Text Coherence. Technical Report CS-07-04, Brown University, Providence, RI, USA.
- Enç, Mürvet. 1991. The semantics of specificity. *Linguistic Inquiry*, 22(1):1–25.
- Evans, Richard. 2001. Applying machine learning toward an automatic classification of it. *Literary and Linguistic Computing*, 16(1):45–57.
- Fabb, Nigel. 1990. The difference between English restrictive and nonrestrictive relative clauses. *Journal of Linguistics*, 26:57–77.
- Féry, Caroline and Kügler, Frank. 2008. Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, 36:680–703.
- Finthammer, Marc and Cramer, Irene. 2008. Exploring and Navigating: Tools for GermaNet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. ELRA, Paris.
- Frege, Gottlob. 1892. Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100(1):25–50.
- Givón, T. 1978. Definiteness and Referentiality. In Greenberg, J., Ferguson, C., and Moravcsik, E., editors, *Universals of Human Language*, pages 291–330. Stanford University Press, Stanford.
- Godfrey, J., Holliman, E., and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP-92*, pages 517–520.
- Goecke, Daniela, Stührenberg, Maik, and Holler, Anke. 2007. Koreferenz, Kospezifikation und Bridging: Annotationsschema. Technical report, Universität Bielefeld/Universität Göttingen. Interne Reports der DFG-Forschergruppe 437 “Texttechnologische Informationsmodellierung”.
- Götze, Michael, Weskott, Thomas, Endriss, Cornelia, Fiedler, Ines, Hinterwimmer, Stefan, Petrova, Svetlana, Schwarz, Anne, Skopeteas, Stavros, and Stoel, Ruben. 2007. Information Structure. In Dipper, Stefanie, Götze, Michael, and Skopeteas, Stavros, editors, *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for*

- Phonology, Morphology, Syntax, Semantics, and Information Structure.*, pages 147–187. Universitätsverlag of Potsdam, Potsdam. ISIS. Working Papers of the SFB 632. Volume 7.
- Gross, Derek, Allen, James F., and Traum, David R. 1993. The Trains 91 Dialogues. Technical report, University of Rochester, Rochester, NY, USA.
- Gundel, Jeanette K., Hedberg, Nancy, and Zacharski, Ron. 1993. Cognitive Status and the Form of Referring Expressions in Discourse. *Language*, 69(2):274–307.
- Gundel, Jeanette K. 1988. *The role of topic and comment in linguistic theory*. Garland Pub., New York.
- Hamp, Birgit and Feldweg, Helmut. 1997. GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL Workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.
- Harabagiu, Sanda M., Bunescu, Răzvan C., and Maiorano, Steven J. 2001. Text and Knowledge Mining for Coreference Resolution. In *Proceedings of NAACL-01*, pages 55–62, Pittsburgh, PA.
- Hasse, Helmut. 1963. *Zahlentheorie*. Akademie-Verlag, Berlin, 2 edition.
- Heeman, Peter and Allen, James F. 1995. The Trains 93 Dialogues. Technical report, University of Rochester, Rochester, NY, USA.
- Heim, Irene and Kratzer, Angelika. 1998. *Semantics in generative grammar*. Blackwell textbooks in linguistics. Blackwell publishers, Cambridge (Mass.), Oxford.
- Heim, Irene. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst. SFB-Papier 73 University of Konstanz.
- Hemforth, Barbara, Konieczny, Lars, and Scheepers, Christoph. 2000. Syntactic attachment and anaphor resolution: The two sides of relative clause attachment. In Crocker, Matthew W., Pickering, Martin J., and Clifton, Charles, editors, *Architectures and mechanisms for language processing*, pages 259–281. Cambridge University Press, Cambridge.
- Hempelmann, Christian F., Dufty, David, McCarthy, Philip M., Graesser, Arthur C., Cai, Zhiqiang, and McNamara, Danielle S. 2005. Using LSA to automatically identify givenness and newness of noun-phrases in written discourse. In Bara, Bruno, Barsalou, Larry, and Bucciarelli, Monica, editors, *Proceedings of the 27th Annual Meetings of the Cognitive Science Society*, pages 941–946, Mahwah, NJ: Erlbaum.
- Herbelot, Aurélie and Copestake, Anne. 2008. Annotating Genericity: How Do Humans Decide? (A Case Study in Ontology Extraction). *The Fruits of Empirical Linguistics*, 1.
- Herbelot, Aurélie. 2011. *Underspecified quantification*. Ph.D. thesis, University of Cambridge, U.K.
- Hirschberg, Julia and Terken, Jacques M. B. 1993. Deaccentuation and persistence of grammatical function and surface position. In *EUROSPEECH*. ISCA.
- Hirschman, Lynette and Chinchor, Nancy. 1997. MUC-7 coreference task definition (ver-

- sion 3.0). In *Proceedings of the Seventh Message Understanding Conference*, Fairfax, Virginia. Science Applications International Corporation.
- Hirschman, Lynette, Robinson, Patricia, Burger, John D., and Vilain, Marc B. 1998. Automating Coreference: The Role of Annotated Training Data. Technical report, AAAI.
- Hiyakumoto, Laurie, Prevost, Scott, and Cassell, Justine. 1997. Semantic and Discourse Information for Text-to-Speech Intonation. In *Proc. ACL Workshop on Concept-to-Speech Generation*.
- Hobbs, Jerry. 1976. Pronoun Resolution. Technical report, Research Report 76-1, New York: Department of Computer Science, City University of New York.
- Hobbs, Jerry R. 1993. The generic information extraction system. In *Proceedings of the 5th conference on Message understanding, MUC5 '93*, pages 87–91, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Hovy, Eduard, Marcus, Mitchell, Palmer, Martha, Ramshaw, Lance, and Weischedel, Ralf. 2006. OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.
- Hunt, Earl B., Martin, Janet, and Stone, Philip T. 1966. *Experiments in Induction*. Academic Press, New York.
- Joachims, Thorsten. 1999. Making large-Scale SVM Learning Practical. In Schölkopf, B., Burges, C., and Smola, A., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.
- Kabadjov, Mijail Alexandrov. 2007. *A Comprehensive Evaluation of Anaphora Resolution and Discourse-new Classification*. Ph.D. thesis, University of Essex, U.K.
- Kamp, Hans and Reyle, Uwe. 1993. *From Discourse to Logic. Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.
- Karttunen, Lauri. 1968. *What do referential indices refer to?* Rand Corporation, Santa Monica, California.
- Karttunen, Lauri. 1969a. Discourse Referents. International Conference on Computational Linguistics.
- Karttunen, Lauri. 1969b. Pronouns and variables. In Binnik, Robert I., Green, Georgia M., Morgan, Jerry L., and Davidson, Alice, editors, *Papers from the Fifth Regional Meeting of the Chicago Linguistic Society*, pages 108–116. University of Chicago, USA, Chicago.
- Karttunen, Lauri. 1974. Presupposition and Linguistic Context. *Theoretical Linguistics*, 1:181–193.
- Kennedy, Christopher and Boguraev, Branimir. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th conference on Computational linguistics - Volume 1, COLING '96*, pages 113–118, Stroudsburg, PA,

- USA. Association for Computational Linguistics.
- Kibble, Rodger and van Deemter, Kees. 2000. Coreference Annotation: Whither? In *In Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1281–1286.
- Kim, Jong-Sun and Evens, Martha W. 1996. Efficient Coreference Resolution for Proper Names in the Wall Street Journal Text. In *MAICS 96*, Bloomington.
- King, Ross D., Feng, Cao, and Sutherland, Alistair. 1995. STALOG: Comparison of classification algorithms on large real-world problems. *Applied Artificial Intelligence*, 9(3):289–333.
- Klinger, Roman and Tomanek, Katrin. 2007. Classical Probabilistic Models and Conditional Random Fields. Technical report, Dortmund University of Technology. Algorithm Engineering Report TR07-2-013.
- Kolb, Peter. 2008. DISCO: A Multilingual Database of Distributionally Similar Words. In et al., Angelika Storrer, editor, *KONVENS 2008 - Ergänzungsband: Textressourcen und lexikalisches Wissen*, Berlin.
- Kotsiantis, Sotiris B. 2007. Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31:249–268.
- Krasavina, Olga and Chiarcos, Christian. 2007. PoCoS: Potsdam Coreference Scheme. In *Proceedings of the First Linguistic Annotation Workshop (LAW). Held in conjunction with ACL-2007*, pages 156–163, Prague.
- Krifka, Manfred, Pelletier, Francis J., Carlson, Gregory N., ter Meulen, Alice, Chiercha, Gennaro, and Link, Godehard. 1995. Genericity: An Introduction. In Carlson, Gregory N. and Pelletier, Francis J., editors, *The Generic Book*, pages 1–124. University of Chicago Press, Chicago.
- Krifka, Manfred. 2004. Bare NPs: Kind-referring, Indefinites, Both, or Neither? In Young, R. B. and Zhou, Y., editors, *Proceedings of Semantics and Linguistic Theory (SALT) XIII*, University of Washington, Seattle. CLC Publications, Cornell.
- Krifka, Manfred. 2008. Basic Notions of Information Structure. *Acta Linguistica Hungarica*, 55:243–276.
- Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA.
- Lafferty, John, McCallum, Andrew, and Pereira, Fernando C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289, San Francisco, CA. Morgan Kaufmann.
- Lawler, John. 1972. Generic to a Fault. *CLS*, 8:247–258.
- Lesgold, Alan, Lajole, Susanne P., Bunzo, Marilyn, and Eggan, Gary. 1992. Sherlock: A Coached Practice Environment for an Electronics Troubleshooting Job. In Larkin, Jill H., Chabay, Ruth W., and Scheftic, Carol, editors, *Computer assisted instruction and intelligent tutoring systems*. Erlbaum, Hillsdale, NJ.

- Lin, Dekang. 1998. An Information-Theoretic Definition of Similarity. In *International Conference on Machine Learning*, pages 296–304.
- Louis, Annie, Joshi, Aravind, Prasad, Rashmi, and Nenkova, Ani. 2010. Using entity features to classify implicit discourse relations. In *Proceedings of the SIGDIAL 2010 Conference*, pages 59–62, Tokyo, Japan, September. Association for Computational Linguistics.
- Lu, Qing and Getoor, Lise. 2003. Link-based classification. In *Proceedings of the 20th International Conference on Machine Learning*, pages 496–503, Washington, D.C.
- Lüdeling, Anke, Ritz, Julia, Stede, Manfred, and Zeldes, Amir. to appear. Corpus Linguistics. In Féry, Caroline and Ishihara, Shinichiro, editors, *The Oxford Handbook of Information Structure*. Oxford University Press, Oxford.
- Ludlow, Peter. 2011. Descriptions. The Stanford Encyclopedia of Philosophy (Winter 2011 Edition).
- Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. In *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP 2005)*, pages 25–32, Vancouver, B.C., Canada.
- Lyons, Christopher. 1999. *Definiteness*. Cambridge University Press.
- Mann, William and Thompson, Sandra. 1987. Rhetorical Structure Theory: A Theory of Text Organization. Technical report, ISI: Information Sciences Institute, Los Angeles, CA. Technical Report ISI/RS-87-190.
- Mann, William and Thompson, Sandra. 1988. Rhetorical Structure Theory: A Theory of Text Organization. *TEXT*, 8.
- Mann, William, Matthiessen, Christian, and Thompson, Sandra. 1993. Rhetorical Structure Theory and Text Analysis. In Mann, William and Thompson, Sandra, editors, *Text Description: Diverse Analyses of a Fund Raising Text*. John Benjamins, Amsterdam.
- Manning, Christopher D. and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.
- Manning, Christopher D., Raghavan, Prabhakar, and Schütze, Hinrich. 2009. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge.
- Marcus, Mitchell P., Santorini, Beatrice, and Marcinkiewicz, Mary Ann. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330. Special Issue on Using Large Corpora.
- Markert, Katja, Hou, Yufang, and Strube, Michael. 2012. Collective Classification for Fine-grained Information Status. In *ACL (1)*, pages 795–804.
- McClosky, David, Charniak, Eugene, and Johnson, Mark. 2006. Reranking and self-training for parser adaptation. In *ACL'06*.
- McNemar, Quinn. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

- Meschkowski, Herbert. 1966. *Mathematisches Begriffswörterbuch*. Bibliographisches Institut AG, Mannheim, 2 edition.
- Meurers, Detmar, Ziai, Ramon, Ott, Niels, and Kopp, Janina. 2011. Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, pages 1–9, Edinburgh, Scotland, UK, July. Association for Computational Linguistics.
- Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.
- Mitkov, Ruslan. 1999. Anaphora Resolution: the state of the art. Unpublished manuscript.
- Moore, Samuel, D’Addario, Daniel, Kurinskas, James, and Weiss, Gary M. 2009. Are Decision Trees Always Greener on the Open (Source) Side of the Fence? In Stahlbock, Robert, Crone, Sven F., and Lessmann, Stefan, editors, *DMIN*, pages 185–188. CSREA Press.
- 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*, pages 49–56.
- Nakatani, Christine H. 1996. Discourse Structural Constraints on Accent in Narrative. In Santen, Jan P. H. Van, Sproat, Richard W., Olive, Joseph P., and Hirschberg, Julia, editors, *Progress in Speech Synthesis*, pages 139–156. Springer, New York.
- Naumann, Karin. 2006. Manual for the Annotation of in-document Referential Relations. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen.
- Nemhauser, George L. and Wolsey, Laurence A. 1988. *Integer and combinatorial optimization*. Wiley.
- Ng, Vincent and Cardie, Claire. 2002. Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002)*, pages 730–736. ACL.
- Ng, Vincent. 2004. Learning noun phrase anaphoricity to improve coreference resolution: issues in representation and optimization. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL ’04, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ng, Vincent. 2009. Graph-Cut-Based Anaphoricity Determination for Coreference Resolution. In *Proceedings of Human Language Technologies 2009: The Conference of the North American Chapter of the Association for Computational Linguistics*, pages 575–583, Boulder, Col.
- Ng, Vincent. 2010. Supervised Noun Phrase Coreference Research: The First Fifteen Years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July. Association for Computational Linguistics.
- Ng, Andrew. 2011. Machine Learning. Online Course, <http://www.ml-class.org>.

- Nissim, Malvina, Dingare, Shipra, Carletta, Jean, and Steedman, Mark. 2004. An Annotation Scheme for Information Status in Dialogue. In *Proceedings of the 4th Language Resources and Evaluation Conference (LREC 2004)*, pages 1023–1026, Lisbon, Portugal.
- Nissim, Malvina. 2003. Annotation Scheme for Information Status in Dialogue. Unpublished manuscript.
- Nissim, Malvina. 2006. Learning Information Status of Discourse Entities. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006)*, pages 94–102, Sydney, Australia.
- Noonan, Harold. 2009. Identity. In Zalta, Edward N., editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA, winter 2009 edition.
- Paice, Chris and Husk, Gareth. 1987. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun *it*. *Computer Speech and Language*, 2:109–132.
- Partee, Barbara, ter Meulen, Alice, and Wall, Robert. 1990. *Mathematical Methods in Linguistics*. Kluwer, Dordrecht.
- Pearson, Karl. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175.
- Poesio, Massimo and Artstein, Ron. 2005. The Reliability of Anaphoric Annotation, Reconsidered: Taking Ambiguity into Account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, CorpusAnno '05, pages 76–83, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Poesio, Massimo and Artstein, Ron. 2008. Anaphoric Annotation in the ARRAU Corpus. In *Proceedings of LREC'2008*.
- Poesio, Massimo and Vieira, Renata. 1998. A corpus-based investigation of definite description use. *Comput. Linguist.*, 24:183–216, June.
- Poesio, Massimo, Bruneseaux, Florence, and Romary, Laurent. 1999. The MATE meta-scheme for coreference in dialogues in multiple languages. In *Proceedings of the ACL Workshop on Standards for Discourse Tagging*, Maryland.
- Poesio, Massimo, Uryupina, Olga, Vieira, Renata, Alexandrov-Kabadjov, Mijail, and Goulart, Rodrigo. 2004. Discourse-new detectors for definite description resolution: a survey and preliminary proposal. In *Proceedings of the ACL Workshop on Reference Resolution at ACL 2004*, Barcelona, Spain, July.
- Poesio, Massimo. 2004a. Discourse Annotation and Semantic Annotation in the GNOME Corpus. In *Proceedings of the ACL Workshop on Discourse Annotation*, Barcelona.
- Poesio, Massimo. 2004b. An empirical investigation of definiteness. In *Proceedings of International Conference on Linguistic Evidence*, Tübingen, Germany.
- Poesio, Massimo. 2004c. The MATE/GNOME Scheme for Anaphoric Annotation, Re-

- visited. In *Proceedings of SIGDIAL*, Boston.
- Postolache, Oana, Kruijff-Korbayová, Ivana, and Kruijff, Geert-Jan M. 2005. Data-driven approaches for information structure identification. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 9–16, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Postolache, Oana. 2005. Learning information structure in the Prague treebank. In *Proceedings of the ACL Student Research Workshop*, pages 115–120, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pradhan, Sameer, Ramshaw, Lance, Marcus, Mitchell, Palmer, Martha, Weischedel, Ralph, and Xue, Nianwen. 2011. CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon.
- Prince, Ellen F. 1981. Toward a Taxonomy of Given-New Information. *Radical Pragmatics*, pages 223–255.
- Prince, Ellen F. 1992. The ZPG Letter: Subjects, Definiteness, and Information-Status. In Mann, William C. and Thompson, Sandra A., editors, *Discourse Description. Diverse linguistic analyses of a fund-raising text*, pages 295–325. Benjamins, Amsterdam.
- Quine, Willard Van Orman. 1986. *Philosophy of logic*. Harvard University Press, 2nd edition.
- Quinlan, J. Ross. 1993. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Rahman, Altaf and Ng, Vincent. 2009. Supervised Models for Coreference Resolution. In *Proceedings of EMNLP*, pages 968–977, Singapore.
- Rahman, Altaf and Ng, Vincent. 2011. Learning the information status of noun phrases in spoken dialogues. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1069–2142, Edinburgh, Scotland, U.K.
- Rashad, M. Z., El-Bakry, Hazem M., Isma’il, Islam R., and Mastorakis, Nikos. 2010. An Overview of Text-To-Speech Synthesis Techniques. In Mastorakis, N., Mladenov, V., and Bojkovic, Z., editors, *LATEST TRENDS on COMMUNICATIONS and INFORMATION TECHNOLOGY*, pages 84–89. WSEAS Press.
- Recasens, Marta, Hovy, Eduard H., and Martí, Maria Antònia. 2010. A typology of near-identity relations for coreference (nident). In *LREC*, pages 149–156.
- Reinhardt, Fritz and Soeder, Heinrich. 1974. *dtv-Atlas zur Mathematik*. Deutscher Taschenbuch Verlag GmbH & Co. KG, München, 6th edition (1984) edition.
- Reinhart, Tanya M. 1976. *The Syntactic Domain of Anaphora*. Ph.D. thesis, Massachusetts Institute of Technology. Dept. of Foreign Literatures and Linguistics.
- Reinhart, Tanya. 1982. Pragmatics and Linguistics: An Analysis of Sentence Topics. *Philosophica*, 27:53–94.
- Reitter, David. 2003. Simple Signals for Complex Rhetorics: On Rhetorical Analysis with

- Rich-Feature Support Vector Models. In *Proceedings of LDV Forum*, pages 38–52.
- Riester, Arndt, Killmann, Lorena, Lorenz, David, and Portz, Melanie. 2007. Richtlinien zur Annotation von Gegebenheit und Kontrast in Projekt A1. Draft version, November 2007. ms.
- Riester, Arndt, Lorenz, David, and Seemann, Nina. 2010. A recursive annotation scheme for referential information status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*, pages 717–722, La Valletta.
- Riester, Arndt. 2009. Partial Accommodation and Activation in Definites. In *Proceedings of the 18th International Congress of Linguists (CIL), Session on Information Structure*, pages 134–152, Seoul. Linguistic Society of Korea.
- Ritz, Julia, Dipper, Stefanie, and Götze, Michael. 2008. Annotation of Information Structure: an Evaluation across different Types of Texts. In Chair), Nicoletta Calzolari (Conference, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios, and Tapias, Daniel, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May. European Language Resources Association (ELRA). <http://www.lrec-conf.org/proceedings/lrec2008/>.
- Ritz, Julia. 2010. Using tf-idf-related Measures for Determining the Anaphoricity of Noun Phrases. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, September.
- Rohrer, Christian and Forst, Martin. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *Proceedings of the Language Resources and Evaluation Conference (LREC-2006)*, Genoa, Italy.
- Rooth, Mats. 1995. Indefinites, Adverbs of Quantification and Focus Semantics. In Carlson, Gregory N. and Pelletier, Francis Jeffry, editors, *The generic book*, pages 265–299. University of Chicago Press.
- Sag, Ivan A. and Pollard, Carl. 1991. An Integrated Theory of Complement Control. *Language*, 67(1):63–113.
- Schiller, Anne, Teufel, Simone, Stöckert, Christine, and Thielen, Christine. 1999. Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Technical report, University of Stuttgart and University of Tübingen.
- Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK. Extended version available at <http://www.ims.uni-stuttgart.de/ftp/pub/corpora/tree-tagger1.pdf>.
- Schwarz-Friese, Monika, Consten, Manfred, and Knees, Mareile, editors. 2007. *Anaphors in text cognitive, formal and applied approaches to anaphoric reference*. Philadelphia : J. Benjamins Pub. Co., Amsterdam. E-BRARY Electronic Book Collection.
- Schwarzschild, R. 1999. GIVENness, AvoidF and other constraints of the placement of accent. *Natural Language Semantics*, 7(2):141–177.
- Soames, Scott. 2003. *Philosophical Analysis In The Twentieth Century: The Age Of Meaning*. Princeton University Press.

- Soon, Wee Meng, Ng, Hwee Tou, and Lim, Daniel Chung Yong. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, 27(4):521–544.
- Stalnaker, Robert. 1974. Pragmatic Presuppositions. In Munitz, Milton K. and Unger, Peter K., editors, *Semantics and Philosophy*, pages 197–213. New York University Press, New York. Reprinted in Davis (1991) and in Stalnaker (1999).
- Stede, Manfred. 2004. The Potsdam Commentary Corpus. In *Association for Computational Linguistics (ACL) 2004 Workshop on Discourse Annotation*, pages 96–102, Barcelona, Spain.
- Stevens, Catherine, Lees, Nicole, Vonwiller, Julie, and Burnham, Denis. 2005. On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference. *Computer Speech and Language*, 19(2):129–146.
- Strawson, Peter F. 1950. On Referring. *Mind*, 59(235):320–344.
- Telljohann, Heike, Hinrichs, Ehrhard W., and Kübler, Sandra. 2003. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Technical report, University of Tübingen.
- Telljohann, Heike, Hinrichs, Ehrhard W., Kübler, Sandra, and Zinsmeister, Heike. 2006. Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Revised Version. Technical report, Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen.
- Tye, Michael. 1991. *The Imagery Debate*. Cambridge: MIT Press.
- Umbach, Carla. 2003. Anaphoric Restriction of Alternative Sets: On the Role of Bridging Antecedents. In *Proceedings of “Sinn und Bedeutung” 7, Konstanz linguistics working papers. No 114*, pages 310–323.
- Uryupina, Olga and Poesio, Massimo. 2012. Domain-specific vs. Uniform Modeling for Coreference Resolution. In et al., Nicoletta Calzolari, editor, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).
- Uryupina, Olga. 2003. High-precision Identification of Discourse New and Unique Noun Phrases. In *Proceedings of the ACL Student Workshop*, pages 80–86, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Uryupina, Olga. 2009. Detecting Anaphoricity and Antecedenthood for Coreference Resolution. In *Procesamiento del lenguaje natural*, volume 42, pages 113–120. Sociedad Española para el Procesamiento del Lenguaje Natural.
- van Deemter, Kees and Kibble, Rodger. 2000. On Coreferring: Coreference in MUC and Related Annotation Schemes. *Computational Linguistics*, 26(4):629–637.
- Vapnik, Vladimir N. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA.
- Versley, Yannick, Moschitti, Alessandro, Poesio, Massimo, and Yang, Xiaofeng. 2008a.

- Coreference systems based on kernels methods. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 961–968.
- Versley, Yannik, Ponzetto, Simone, Poesio, Massimo, Eidelman, Vladimir, Jern, Alan, Smith, Jason, Yang, Xiaofeng, and Moschitti, Alessandro. 2008b. BART: A Modular Toolkit for Coreference Resolution. In *Companion Volume of the Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL 2008)*.
- Versley, Yannick. 2006. A constraint-based approach to noun phrase coreference resolution in German newspaper text. In *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS 2006)*.
- Vieira, Renata and Poesio, Massimo. 2000. Corpus-based Development and Evaluation of a System for Processing Definite Descriptions . In *Proceedings of the 18th conference on Computational linguistics - Volume 2, COLING '00*, pages 899–903, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Vilain, Marc, Burger, John, Aberdeen, John, Connolly, Dennis, and Hirschman, Lynette. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages=45–52.
- von Heusinger, Klaus. 2002. Specificity and Definiteness in Sentence and Discourse Structure. *Journal of Semantics*, pages 245–274.
- Walker, Christopher, Strassel, Stephanie, Medero, Julie, and Maeda, Kazuaki. 2006. ACE 2005 Multilingual Training Corpus.
- Washburn, Margaret. 1898. The Psychology of Deductive Logic. *Mind*, 7(28):523–530.
- Webber, Bonnie L. 1988. Discourse Deixis and Discourse Processing. Technical report, Department of Computer and Information Science, University of Pennsylvania. Technical Report MS-CIS-88-75.
- Weischedel, Ralph, Pradhan, Sameer, Ramshaw, Lance, Micciulla, Linnea, Palmer, Martha, Xue, Nianwen, Marcus, Mitchell, Taylor, Ann, Babko-Malaya, Olga, Hovy, Eduard, Belvin, Robert, and Houston, Ann. 2007. OntoNotes Release 1.0. Technical report, Linguistic Data Consortium, Philadelphia.
- Wilks, Yorick. 1973. Preference Semantics. Technical report, Stanford AI Laboratory memo AIM-206, Stanford University.
- Winograd, Terry. 1972. *Understanding natural language*. Academic Press, New York.
- Witten, Ian H. and Frank, Eibe. 2005. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufman, San Francisco, 2 edition.
- Yates, Frank. 1934. Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235.
- Zeldes, Amir, Ritz, Julia, Lüdeling, Anke, and Chiarcos, Christian. 2009. ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In *Proceedings of Corpus Linguistics 2009*, Liverpool, UK.
- Zhou, Guo-Dong and Kong, Fang. 2011. Learning noun phrase anaphoricity in coreference resolution via label propagation. *Journal of Computer Science and Technology*,

26(1):34–44.

Zhu, Xiaojin and Ghahramani, Zoubin. 2002. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, CMU.

Index

- κ , 75
- abstract concepts, 51, 70, 72
- accessible, 49
- ACE, 139
- aggregation, 59
- anaphoric expressions, 15
- apposition, 51, 70, 72
- ARRAU, 5, 54, 55, 101, 123
- asserted identity, 51, 70
- asymmetry
 - specific vs. generic, 32
- Büring, 39
- bias, 90
- binary, xv
- bindee, 40
- binder, 40
- binding, 39, 51, 70, 72, 134
 - semantic, 39
 - syntactic, 39
 - binding theory, 39
- bridging, 48, 51, 70, 72
- c-command, 4, 39
- C4.5 decision trees, 88
- cataphor, 69, 70, 72
- classification functions, 90
- clause, 64, 66
- clauses as referents, 62
- cognitive status, 4
- Common Ground, 14
- comparison of classification techniques, 92
- conditional random field, 92
- conjunction, 64, 66
- context, 57, 74
- coreference, 57, 69
- coreference between indefinites, 32, 126
- coreference resolution, 2, 44, 140
- coreference set, 75
- coreferent, 3
- cost function, 90
- CRF, 92
- decision trees, 88
- definiteness, 51, 68, 70, 72
- DIRNDL, 5, 54, 57, 82
- discontinuous markable, 62, 64, 66
- discontinuous markable, 59, 78
- discourse-given, 3
- discourse-givenness, 69
- DRS, 15
 - examples, 15
- DRT, 13, 15, 25, 26
- dyadic operator, 31
- embedded expression, 61, 64, 66
- embedding, 61
- equivalence relation, 35
- evaluation
 - measures for annotation evaluation, 76
- evaluation of algorithms, 93
- evaluation of annotation, 57, 75
- event anaphor, 46, 51, 70, 72
- event anaphors, 134
- expression
 - form of an expression, 15
- extension, 14, 33
- f-measure, 75
- false negative, 93
- false positive, 93
- feature coding
 - boolean, 83
 - categorial, 83
 - numeric, 83
- formalization, 74
- formalization of annotation task, 57
- function value, 51, 70, 72
- fusion, 62, 64, 66

genericity, 37
 GermaNet, 87
 gerund, 64, 66
 givenness, 15
 givenness hierarchy, 49
 graph
 directed, 75
 Gundel et al. (1993), 4, 49

 hearer, 14

 ICA, 92
 ID, 75
 identity, 35
 illocutionary act, 19
 ILP, 91
 inferable, 49
 information status, 4, 69
 information structure, 7
 integer linear programming, 91
 intension, 14, 33
 inter-annotator agreement, 75
 interrogative constituent, 61, 64, 66
 iterative collective classification, 92

 J48 decision trees, 88, 110
 junctors, 16

 kappa, 75
 kernels, 90
 kinds
 reference to, 29, 51, 70, 72
 Kotsiantis (2007), 92
 Krifka's cat, 15

 label propagation, 92
 locational expression, 61, 64, 66
 logarithmic scale, 114
 logical operators, 16
 logistic regression, 90

 markable, 57, 64
 markable boundaries, 62
 McNemar's χ^2 , 95
 measures for annotation evaluation, 76
 mediated, 49
 metonym, 37
 metonymy, 123

 monadic operator, 31
 MUC-6, 5, 82
 MUC-7, 5, 54, 82, 101, 116, 136
 multiple heads, 62

 near-identity, 38
 negation, 29, 30
 new, 4, 7, 70, 72
 non-referring, 41, 51, 70, 72
 anaphor to, 41
 notation
 index, 3
 subscript, 3
 underline, 3
 null hypothesis, 94
 numeric expression, 61, 64, 66

 OntoNotes, 5, 54, 55, 82, 101, 110, 136
 overfitting, 90

 paragraph, 64, 66
 paragraphs as referents, 62
 paycheck-example (Karttunen), 38
 PCC (Potsdam Commentary Corpus), 5, 54, 56, 101
 Pearson's χ^2 , 95
 percent agreement, 75
 possessive pronoun, 64, 66
 possessive *s*, 62, 64, 66
 precision, 75
 predication, 51, 70, 72
 present participle, 64, 66
 Prince (1981, 1992), 4, 49
 pronoun
 possessive, 59, 64, 66
 reciprocal, 61, 64, 66
 reflexive, 61, 64, 66
 relative, 60, 64, 66
 property, 15
 proposition anaphor, 51, 70, 72
 prosody, 6

 quantification, 29
 quantifier, 16

 random split, 110
 recall, 75
 reciprocal pronoun, 64, 66

referent, 14, 57
 referentiality, 57, 59, 68
 referring expression, 14
 referring expressions, 2
 reflexive pronoun, 64, 66
 reflexivity, 35
 relation, 15
 unary, 17
 relative pronoun, 64, 66
 Ripper rule learner, 89, 110
 rule learners, 89

 scope, 16, 30, 31, 33
 skewed distribution, 93
 speaker, 14
 specificity, 51, 68, 70, 72
 in OntoNotes, 68
 subscript index, 3
 summation, 59
 support vector machines, 90, 110
 SVM, 90, 110
 Switchboard corpus, 54, 55, 82
 symmetry, 35, 40

 TüBa-D/Z, 5, 54, 56, 101, 126
 temporal expression, 61, 64, 66
 titles, 64, 66
 titles in markables, 62
 trace, 61, 64, 66
 transitivity, 35
 true negative, 93
 true positive, 93
 type/token, 40

 unary, xv, 17
 underline, 3
 universal quantifier, 16

 variable, 15
 variance, 90
 verb, 64, 66
 verbs as referents, 62

 weighted average, 94
 word sense, 55
 WordNet, 55, 87

 Yates' correction, 95

 zero form, 61, 64, 66