Schwierigkeiten in der automatischen Verarbeitung von Anaphora

Lizenziatsarbeit der Philosophischen Fakultät
der Universität Zürich

Renata Sevcikova
Matrikelnummer 98-717-259

Angefertigt am
Institut für Computerlinguistik
der Universität Zürich
Prof. Dr. Michael Hess

Abgabe der Arbeit:
Dezember 2005

Zusammenfassung

Anaphernresolution (AR) nimmt eine Schlüsselrolle in vielen Anwendungsgebieten der automatischen Sprachverarbeitung ein und ist daher ein beliebtes Forschungsgebiet. Obwohl in den letzten zwanzig Jahren vielversprechende Arbeiten vorgestellt wurden, bleiben viele Probleme in der automatischen Behandlung von Anaphern noch zu lösen. Diese Arbeit konzentriert sich vorwiegend auf die Schwierigkeiten bei der automatischen Auflösung von Proformen und vollständigen Nominalphrasen und versucht auch Lösungsansätze aufzuzeigen. Weiter werden Schwierigkeiten bei der Auswahl des korrekten Antezedens aus einer Liste von möglichen Kandidaten behandelt. Der Schwerpunkt liegt hierbei auf Diskrepanzen im Numerus, Genus und der Belebtheit zwischen Anapher und Antezedens. Insbesondere dort, wo Weltwissen zur Auflösung von Anaphern benötigt wird, stösst man in der automatischen Verarbeitung an Grenzen.

Abstract

Anaphora Resolution (AR) is a key task in many applications of natural language processing. AR attracted the attention of many researches and in the last 20 years much promising work on the topic has been reported. Nevertheless, AR remains a daunting task with a lot of problems to deal with. This paper concentrates on the problems in the resolution of proforms and full noun phrases and tries to present some solutions. Furthermore, difficulties in the resolution process are discussed. The emphasis lies on number, gender and animacy mismatches between anaphor and antecedent. Especially the lack of semantic, domain and world knowledge imposes serious limitations on AR.

Inhaltsverzeichnis

ZUSAMMENFASSUNG ..... **I**

ABSTRACT ..... **I**

1 EINLEITUNG **1**

1.1 PROBLEMSTELLUNG DER ARBEIT 1

1.2 ZIELSETZUNG UND AUFBAU DER ARBEIT 2

2 DAS PHÄNOMEN ANAPHORA **4**

2.1 LINGUISTISCHE GRUNDLAGEN 4

2.2 COMPUTERLINGUISTISCHE ANSÄTZE 5

3 VERSCHIEDENE TYPEN VON ANAPHERN **10**

3.1 UNTERSCHEIDUNG NACH SYNTAKTISCHER REALISATION 10

3.2 UNTERSCHEIDUNG NACH LAGE, REFERENZVERMÖGEN UND ERSCHLIESSBARKEIT 11

4 SCHWIERIGKEITEN MIT DEN VERSCHIEDENEN TYPEN VON ANAPHERN **13**

4.1 NICHT-LEXIKALISCHE PROFORMEN UND IHRE AUTOMATISCHE BEHANDLUNG 13

4.1.1 Erkennen und Ausfiltern von nicht-anaphorischen Proformen *14*

4.1.1.1 Ausfiltern von kataphorischen, deiktischen und generischen Pronomen 14

4.1.1.2 Ausfiltern von pleonastischen Pronomen 16

4.1.1.2.1 Erkennung von pleonastischen Pronomen mit Suchmustern 18

4.1.1.2.2 Erkennung von pleonastischen Pronomen mit ML 24

4.1.2 Erkennen von anaphorischen Proformen *25*

4.1.2.1 Automatische Auflösung von *One*-Anaphern 26

4.1.2.1.1 Auflösung von *One*-Anaphern mit Ähnlichkeitskriterien und semantischen Hinweisen 26

4.1.2.1.2 Auflösung von *One*-Anaphern mit ML 27

4.1.2.2 Automatische Behandlung von Ereignis-Anaphern 29

4.1.2.3 Automatische Behandlung von assoziativen Pronomen 31

4.1.3 Pronomenresolutionssysteme *32*

4.1.3.1 Pronomenresolutionssysteme mit syntaktischen und semantischen Merkmalen 32

4.1.3.1.1 Das System von [Lappin et al. 1994] 32

4.1.3.1.2 Das System von [Denber 1998] 36

4.1.3.1.3 Das System von [Mitkov 2002a] 36

4.1.3.1.4 Das System von [Siddharthan 2003] 41

4.1.3.2 Pronomenresolutionssysteme mit Techniken des ML 44

4.1.3.2.1 Das System von [Kehler et al. 2004] 44

4.2 VOLLSTÄNDIGE NPS UND IHRE AUTOMATISCHE BEHANDLUNG 45

4.2.1 Erkennen und Ausfiltern von nicht-anaphorischen NPs *45*
4.2.1.1 Erkennen und Ausfiltern von exophorisch verweisenden vollständigen NPs 45
4.2.1.2 Erkennen und Ausfiltern von diskursneuen definiten NPs 47
4.2.1.3 Algorithmus zur Erkennung von nicht-anaphorischen NPs 49
4.2.2 Automatische Auflösung von anaphorischen NPs *51*
4.2.2.1 Automatische Auflösung von Identitätsanaphern 51
4.2.2.2 Automatische Auflösung von assoziativen Anaphern 54
4.2.2.2.1 Auflösung von assoziativen Anaphern mit Heuristiken und WordNet 56
4.2.2.2.2 Auflösung von assoziativen Anaphern mit assoziativen Axiomen aus Korpora und WordNet 57
4.2.2.2.3 Auflösung von assoziativen Anaphern mit fokusbasierten Techniken 59
4.2.2.2.4 Auflösung von assoziativen Anaphern mit Hilfe des Webs 60
4.2.2.2.5 Auflösung von assoziativen Anaphern mit Ähnlichkeitskriterien 63
4.2.2.2.6 Auflösung von assoziativen Anaphern mit vektorbasierten Methoden 64
4.2.2.2.7 Auflösung von assoziativen Anaphern mit Techniken des ML 64
4.2.2.3 Automatische Auflösung von *Other*-Anaphern 67
4.2.2.4 Ein Algorithmus zur Erkennung von anaphorischen NPs 70
4.2.3 AR-Systeme zur Behandlung von vollständigen NPs *71*
4.2.3.1 Der Ansatz von [Vieira et al. 2000] 71
4.2.3.2 Der Ansatz von [Postolache 2004] 73

5 DISKREPANZEN IN DEN KONGRUENZMERKMALEN 80
5.1 DISKREPANZEN IM NUMERUS 83
5.1.1 Numerusdiskrepanzen aufgrund komplexer Antezedenten *84*
5.1.2 Numerusdiskrepanzen aufgrund kollektiver Nomen *86*
5.1.3 Numerusdiskrepanzen aufgrund von Gruppenverkörperungen *87*
5.1.4 Numerusdiskrepanzen aufgrund von Genusunterspezifikation *87*
5.1.5 Numerusdiskrepanzen aufgrund assoziativer Anaphern *88*
5.2 DISKREPANZEN IM GENUS 89
5.2.1 Genusdiskrepanzen aufgrund von Anomalien *89*
5.2.2 Genusdiskrepanzen aufgrund von Eigennamen *89*
5.3 BELEBTHEITSINFORMATION ZUR AUFLÖSUNG VON ANAPHERN 91
5.3.1 Algorithmus zur Bestimmung der Belebtheitsinformation *92*

6 FAZIT 95

7 LITERATURVERZEICHNIS VI

8 GLOSSAR XIII

LEBENSLAUF XVII

Abbildungsverzeichnis

Abb. 1 Darstellung von Koreferenz, basierend auf [Linke et al. 2004:246] 4
Abb. 2 Darstellung von referentiellen Ausdrücken von [Halliday et al. 1976:35] 5
Abb. 3 Darstellung der Komponenten von AR-Systemen, basierend auf [Cristea 2001:26] 7
Abb. 4 Messwerte für die Evaluation von AR-Systemen, basierend auf [Byron 2004a:42] und [Jurafsky et al. 2000:578] 8
Abb. 5 *Nicht-triviale* und *kritische Erfolgsrate*, basierend auf [Mitkov 2001:257] 9
Abb. 6 Eigene Darstellung der Unterscheidungsarten von Anaphern 10
Abb. 7 Fehlerquellen in der Pronomenresolution, basierend auf [Evans 2002:55] 13
Abb. 8 Suchmuster für *it..to*- und *it..that*-Konstrukte von [Paice et al. 1987:112,114,115] 19
Abb. 9 Suchmuster für *it..which* und *it..who*-Konstrukte von [Paice et al. 1987:117,118] 20
Abb. 10 Mögliche Konstruktionen, in denen das Pronomen *it* pleonastisch ist [Lappin et al. 1994:538] 21
Abb. 11 Modale Adjektive und kognitive Verben zur Erkennung von pleonastischen Pronomen [Lappin et al. 1994:538] 21
Abb. 12 Suchmuster zur Erkennung von pleonastischen Pronomen von [Denber 1998:20-22] 21
Abb. 13 Modellierung in der LINK-Grammatik, basierend auf [Strüber 2001:7] 23
Abb. 14 Merkmale von [Litrán et al. 2004:3] zur Erkennung von pleonastischen Pronomen 25
Abb. 15 Auflösung eines Demonstrativpronomens von [Byron 2002:85] 31
Abb. 16 Syntaktische Einschränkungen zur Ausschliessung von Koreferenz, basierend auf [Lappin et al. 1994:538] 33
Abb. 17 Bedingungen an Antezedenten von Reflexivpronomen und reziproken Pronomen [Lappin et al. 1994:538,539] 33
Abb. 18 Beispiel einer Liste mit den Verben und den vorangehenden und nachfolgenden Köpfen der NPs [Mitkov et al. 2002b:5] 39
Abb. 19 Verbesserung der Erfolgsrate durch Anwendung eines bilingualen AR-Systems von [Mitkov 2002a:209] 41
Abb. 20 Suchmuster für grammatikalische Funktionen von [Siddharthan 2003:8] 42
Abb. 21 Kongruenzmerkmale von [Siddharthan 2003:9] 42
Abb. 22 Kongruenzmerkmale für Relativpronomen, basierend auf [Siddharthan 2003:10] 43
Abb. 23 Suchmuster zur Erkennung von Postmodifikation, die auf diskursneue Information hindeutet, von [Vieira et al. 2000:555] 48
Abb. 24 Algorithmus zur Generierung von Hauptnomen-Mustern zur Erkennung von nicht-anaphorischen definiten NPs von [Bean et al. 1999:376] 49
Abb. 25 Algorithmus zur Erkennung von nicht-anaphorischen definiten NPs, basierend auf [Bean et al. 1999:377] 50
Abb. 26 Axiome und Verallgemeinerungen von [Meyer et al. 2002:4] 59
Abb. 27 Suchmuster für assoziative Anaphern von [Bunescu 2003:3] 61
Abb. 28 Berechnung des Assoziationswertes von [Bunescu 2003:3] 61
Abb. 29 Berechnung des Bezugswortes in Meronymie-Beziehungen von [Poesio et al. 2002:1223] 64

Abb. 30 Distanz zwischen assoziativen Anaphern und ihrem Antezedens basierend auf [Poesio 2003:5] 65
Abb. 31 Google-Distanz von [Poesio et al. 2004:4] 65
Abb. 32 Kürzeste WordNet-Distanz von [Poesio et al. 2004:5] 66
Abb. 33 WordNet-Distanz von [Poesio et al. 2004:5] 66
Abb. 34 Suchmuster für *Other*-Anaphern im Web-Ansatz von [Markert et al. 2003:5] 68
Abb. 35 Häufigkeitsberechnung der Vorkommen von Anapher-Antezedens-Paaren in Google von [Markert et al. 2003:5] 68
Abb. 36 Auflistung von korrekt und falsch aufgelösten *Other*-Anaphern im webbasierten Ansatz von [Markert et al. 2003] und dem WordNet-Ansatz LEX, basierend auf [Markert et al. 2003:5] 69
Abb. 37 Suchmuster für assoziative Anaphern in Meronymie-Beziehungen von [Markert et al. 2003:6] 70
Abb. 38 Darstellung des Systems von [Vieira et al. 2000:545] zur Behandlung von definiten NPs 71
Abb. 39 Entscheidungsbaum zur Klassifikation von definiten NPs von [Vieira et al. 2000:577] 72
Abb. 40 Zwei Evaluationsversionen des AR-Systems von [Vieira et al. 2000:581] 72
Abb. 41 Darstellung einer Koreferenz-Relation im RARE-System, basierend auf [Postolache 2004:14] 74
Abb. 42 Darstellung einer assoziativen Relation im RARE-System, basierend auf [Cristea et al. 2005:20] 74
Abb. 43 Darstellung einer anaphorischen Relation im RARE-System, basierend auf [Postolache 2004:14] 74
Abb. 44 Syntaxbaum für die Sätze in Beispiel (80) 81
Abb. 45 Code zur Ausfilterung von indefiniten NPs, die nicht als Antezedens zugänglich sind [Byron et al. 2004b:22] 82
Abb. 46 Numerusinkongruenz zwischen Pluralanpahern und ihrem Antezedens, basierend auf [Barbu et al. 2002:3] 84
Abb. 47 Formeln für die Berechnung von Belebtheitsinformationen von [Evans et al. 2000:156] 93
Abb. 48 Algorithmus von [Evans et al. 2000:160] zur Ermittlung von Belebtheitsinformationen 93

Abkürzungsverzeichnis

AR	Automatische Anaphernresolution
ID	Identifikationsnummer
ML	Maschinenlernen/Techniken des Maschinenlernens
NP	Nominalphrase
POS	Part of Speech (Wortart)
PP	Präpositionalphrase
VP	Verbalphrase
Web	World Wide Web (www)

1 Einleitung

1.1 Problemstellung der Arbeit

Das linguistische Phänomen ANAPHORA beschreibt die Verweisung auf bereits erwähnte Einheiten in einem Text. Die rückverweisenden Ausdrücke werden Anaphern genannt; die bereits erwähnte Einheit, auf die sich die Anapher bezieht, wird als Antezedens bezeichnet. Im Beispiel (1) sind es die Pronomen *his* und *he*, welche auf den zuvor erwähnten *Edward Bear* hindeuten.

(1) *HERE is **Edward Bear**, coming downstairs now, bump, bump, bump, on the back of **his** head, behind Christopher Robin. It is, as far as **he** knows, the only way of coming downstairs, but sometimes **he** feels that there really is another way, if only **he** could stop bumping for a moment and think of it.* [1]

Anaphernresolution kann als Prozess der Bestimmung des Antezedens einer Anapher definiert werden. Der Prozess der automatischen Anaphernresolution, im Folgenden mit AR abgekürzt, kann in drei Phasen gegliedert werden: Die Identifikation der Anapher, das Finden von möglichen Antezedenten und die Auswahl eines Antezedens.

AR ist von grosser Bedeutung in zentralen computerlinguistischen Anwendungsbereichen wie der Informationsextraktion, der Textzusammenfassung und der maschinellen Übersetzung. Aufgrund dieser Schlüsselfunktion wurden viele verschiedene Ansätze zur automatischen Auflösung von Anaphern entwickelt.

Als Veranschaulichung eines Anwendungsbereichs soll hier auf das Bildsuchsystem von [Denber 1998] verwiesen werden, das einen Algorithmus[2] zur AR beinhaltet. Bei der Suchanfrage in (2) verknüpft der AR-Algorithmus das Pronomen *he* mit der Nominalphrase (NP) *a man* und das Pronomen *she* mit der NP *a woman*. Ohne diese Verknüpfung werden sowohl Bilder mit einem stehenden Mann und einer sitzenden Frau als auch Bilder mit einem sitzenden Mann und einer stehenden Frau gefunden. Ausserdem werden auch Bilder gefunden, auf denen sowohl die Frau als auch der Mann steht und solche auf denen beide sitzen.

(2) *A man and a woman. He is standing, she is sitting.*

[Denber 1998:5]

[1] Milne, Alan Alexander. *Winnie-The-Pooh*. [Stand: 22.8.2005] http://www.machaon.ru/pooh/chap1.html

[2] Verfahren mit bestimmten (teilweise sich wiederholenden) Handlungsschritten zur Lösung eines Problems

Wenn nun aber die Pronomen mit einem AR-System aufgelöst werden können, kann die Suchanfrage präzisiert und die Genauigkeit der Suchergebnisse verbessert werden.
Was bereitet denn eigentlich Schwierigkeiten in der AR? [Cristea 2001:15] antwortet „Nothing – everything is so simple!“ und gibt (3) als Beispiel, in dem das Pronomen mit seinem Antezedens in den Kongruenzmerkmalen übereinstimmt.

(3) ***John*** *has just arrived.* ***He*** *seems tired.*

[Cristea 2001:15]

Er fügt aber an, „Nothing indeed, but imagine letting the machine go wrong...“

(4) *If the baby does not thrive on the raw milk, boil* ***it****.*

[Cristea 2001:16]

Die Schwierigkeit der AR besteht also darin, dass dem Computer, im Gegensatz zum Menschen, nur beschränktes Wissen zur Verfügung steht. Zur korrekten Auflösung des Pronomens *it* muss man wissen, dass üblicherweise Milch, und nicht ein Kind, gekocht wird. In einigen Fällen ist auch Kontextwissen zur korrekten Interpretation der Anapher erforderlich. In Beispiel (5) kann sich *his* entweder auf *Jack* oder auf *Joe* beziehen.

(5) *While Jack was talking, Joe finished his beer.*

[Denber 1998:6]

1.2 Zielsetzung und Aufbau der Arbeit

Ziel dieser Arbeit ist es, einen Überblick über die Schwierigkeiten der AR zu geben und aktuelle Lösungsansätze zu den angesprochenen Problemen vorzustellen. Es werden allerdings nicht ganze Systeme detailliert beschrieben, sondern vielmehr nur einzelne Aspekte beleuchtet. Ausserdem beschränkt sich diese Arbeit vorwiegend auf die AR in der englischen Sprache.
Nach der Einleitung werden in Kapitel 2 linguistische Grundlagen zum Phänomen Anaphora vermittelt und in einem kurzen Überblick wird die Entwicklung in der AR erläutert. Ausserdem werden einige Messwerte vorgestellt, die bei der Evaluation von AR-Systemen eine wichtige Rolle spielen. In Kapitel 3 werden die verschiedenen Typen von Anaphern aufgezeigt. In Kapitel 4 wird auf die Probleme und Lösungsansätze in Bezug auf den Typ der Anapher eingegangen. Der Schwerpunkt liegt hierbei auf Anaphern, die als nicht-lexikalische

Proformen[3] oder als vollständige NPs realisiert sind. Die Behandlung von anderen Anapherntypen wie Adverbanaphern, Verbalphrasenanaphern (VP-Anaphern) und Nullanaphern wird aufgrund der eher raren Lösungsansätze nicht näher beleuchtet. Anschliessend werden in Kapitel 5 Diskrepanzen in den Kongruenzmerkmalen zwischen Antezedens und Anapher diskutiert, die den Prozess der AR erschweren. Kapitel 6 greift die wichtigsten Erkenntnisse nochmals auf und gibt einen Ausblick in die Zukunft der AR.

[3] Proformen sind Verweisformen im Text

2 Das Phänomen Anaphora

In diesem Kapitel wird zuerst das linguistische Phänomen Anaphora beschrieben. Dazu werden verwandte Konzepte und wichtige Begriffe im Zusammenhang mit der AR eingeführt. Weiter wird die Entwicklung von AR-Systemen von den Anfängen bis zur Gegenwart kurz beleuchtet und es werden Messwerte für die Evaluation von AR-Systemen eingeführt.

2.1 Linguistische Grundlagen

Unter Anaphernresolution versteht man die Auflösung des Bezugs einer Anapher. Dabei muss das Antezedens vom Referenten unterschieden werden. Während das Antezedens nur eine linguistische Form darstellt, handelt es sich beim Referenten um eine Einheit aus der realen Welt. Neben konkreten Referenzobjekten gehören auch Einheiten wie Konzepte, Vorstellungen und Ideen der realen Welt an.

Ein wichtiges Konzept für Anaphora ist Koreferenz. Von Koreferenz spricht man, wenn ein verweisender Ausdruck und sein Bezugswort den gleichen Referenten in der realen Welt haben. Zwei referentielle Ausdrücke sind koreferent, wenn sie sich auf die gleiche Einheit in der realen Welt beziehen, wie in Abbildung 1 dargestellt ist. Verweisende Ausdrücke können auch koreferentielle Ketten bilden und der Prozess zur Erkennung dieser Ketten wird Koreferenzresolution genannt.

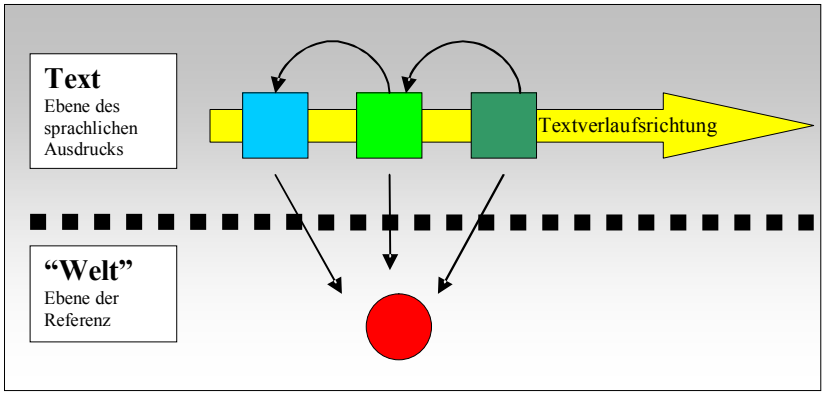

Abb. 1 Darstellung von Koreferenz, basierend auf [Linke et al. 2004:246]

Koreferenz ist aber nicht mit Anaphora gleichzusetzen, denn eine Relation zwischen einer Anapher und ihrem Antezedens kann anaphorisch sein, ohne dass sie koreferentiell ist. Im Beispiel (6) haben die Anapher und das Antezedens kein bestimmtes Referenzobjekt in der realen Welt.

(6) *Every man is entitled to **his** opinion.*

[BNC:BN3 1245]

Umgekehrt können zwei Begriffe auch koreferentiell sein, ohne dass sie eine anaphorische Beziehung eingehen, wie es der Fall mit dokumentübergreifenden Ausdrücken ist. So kann beispielsweise Tony Blair in mehreren Dokumenten erwähnt sein. Diese Erwähnungen stehen aber in keinem textuellen Bezug zueinander.

Abbildung 2 zeigt, welche referentiellen Ausdrücke es neben Anaphern gibt.

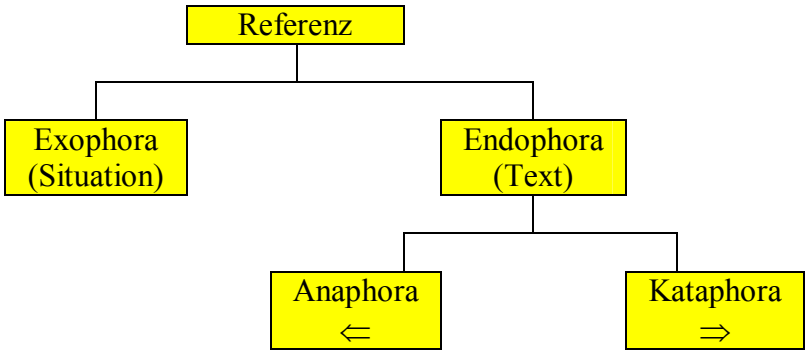

Abb. 2 Darstellung von referentiellen Ausdrücken von [Halliday et al. 1976:35]

Referentielle Ausdrücke können in Exophora und Endophora unterteilt werden. Exophora sind Ausdrücke, die sich nicht auf andere sprachliche Einheiten im Text beziehen, sondern auf textexterne Elemente verweisen. So wird beispielsweise für die Interpretation von *I, now, there, the sun, the president of the U.S.A.* in Dialogen keine weitere sprachliche Einheit benötigt. Endophora sind sprachliche Ausdrücke, die sich auf andere sprachliche Einheiten im Text beziehen. Im Gegensatz zu Exophora werden bei Endophora textinterne Beziehungen hergestellt. Endophora ist der Oberbegriff für Anaphern (rückbezüglich verweisende Ausdrücke) und Kataphern (vorauszeigende Ausdrücke). Beispiele für Anaphern wurden bereits gegeben. Ein Beispiel für eine Katapher ist in (7) aufgeführt.

(7) ***She** is now as famous as her ex-boyfriend. From the deserts of Kazakhstan to the south seas of Tonga, everyone knows **Monica Lewinsky**.*

[Mitkov et al. 2002c:19]

2.2 Computerlinguistische Ansätze

Dieses Kapitel stützt sich auf die Vorarbeit von [Mitkov 2002a:68-125], der die bedeutendsten Systeme in der Geschichte der AR zusammengetragen hat. Für eine ausführlichere Übersicht und Quellenangaben wird daher empfohlen [Mitkov 2002a:68-125] zu konsultieren.

Die ersten AR-Ansätze gehen auf die 60er und 70er Jahre zurück und basieren vorwiegend auf Heuristiken[4] und Mustervergleichen[5]. Beispiele sind das System STUDENT von Bobrow (1964)[6] zur Lösung von einfachen Algebra-Aufgaben und das simple Dialogsystem SHRDLU von Winograd (1972).

Die Arbeiten der 70er und 80er Jahre sind eher theorielastig und haben zum Ziel, möglichst viele Typen von Anaphern aufzulösen. Sie sind auf ein hohes Ausmass von linguistischem und aussersprachlichem Wissen angewiesen und ihre Implementation ist daher stark eingeschränkt, wenn überhaupt realisierbar. Auch müssen diese wissensreichen Ansätze von Menschen überarbeitet werden. Die Bedeutung dieser Arbeiten ist jedoch nicht zu unterschätzen, da sie die Grundsteine für die heutigen Ansätze bilden.

Syntaktische und semantische Beschränkungen, teilweise aber auch Wissen über das Diskursthema und den lokalen Fokus, bilden das Fundament für diese Ansätze. Eine semantische Einschränkung ist beispielsweise, dass in der Phrase „*He drank it*“ [Mitkov 2002a:80] sich das Pronomen *it* nur auf eine Flüssigkeit beziehen kann. Es gibt aber auch Beschränkungen, die auf Weltwissen oder Pragmatik[7] beruhen. In Beispiel (8) muss das Pronomen *he* auf *Martin* verweisen, denn die Bedingung an das Verb *give* ist, dass man nicht im Besitz von etwas Weggegebenem sein kann, und die Bedingung and das Verb *eat* verlangt, dass das Agens das zu essende Objekt hat.

(8) *George gave **Martin** an apple. **He** ate the apple.*

[Mitkov et al. 2002c:88]

Wichtige Arbeiten in dieser Zeit stammen von Woods (1972), Hobbs (1976), Sidner (1979), Webber (1979), Brennan et al. (1987), Walker (1989), Carter (1986), Rich und LuperFoy (1988) und Carbonell und Brown (1988).

In den 90er Jahren wurden die wissensreichen Ansätze durch wissensarme AR-Strategien abgelöst. Man kam von der Komplexität der wissensreichen Ansätze ab und es wurden implementierbare Systeme entwickelt. Diese Ansätze beschränkten sich vorwiegend auf die Behandlung von Anaphern in bestimmten Textgenres. Zudem wurden diese wissensarmen An-

---

[4] „Heuristiken sind Lösungsfindeverfahren, die auf Hypothesen, Analogien oder Erfahrungen aufgebaut sind“ [Balzert:1999:560]. Heuristiken nutzen problemspezifische Informationen, um Lösungen in umfassenden Problemräumen zu finden.

[5] Der englische Begriff lautet Pattern Matching. Beim Mustervergleich werden Zeichenketten auf ihre Ähnlichkeit miteinander verglichen. Die zentrale Frage dabei ist, wie ähnlich zwei Zeichenketten sein müssen, damit sie als gleich gelten.

[6] Quellenangaben in runden Klammern sind nicht im Literaturverzeichnis aufgeführt, aber in [Mitkov 2002a] zu finden

[7] Teildisziplin der Sprachwissenschaft, die sich mit der Verwendung der Sprache befasst. Man kann Sprache verwenden, um beispielsweise jemanden zu bitten oder jemandem zu drohen.

sätze auch durch das Aufkommen von verlässlichen computerlinguistischen Werkzeugen wie POS-Taggern[8] und Parsern[9] sowie Ressourcen wie Korpora,[10] Ontologien[11] und Evaluationsmethoden begünstigt. Ausserdem hat man begonnen, Methoden des Maschinenlernens[12], im Folgenden mit ML abgekürzt, probabilistische Ansätze[13] und mehrsprachige Systeme zur Behandlung von Anaphern einzusetzen.
Dagan und Itai (1990), [Lappin et al. 1994], Kennedy und Boguraev (1996), Baldwin (1997) und [Mitkov 2002] bieten Pronomenresolutionssysteme an. [Vieira et al. 2000], [Bean et al. 1999], [Bunescu 2003], [Cristea et al. 2005], [Markert et al. 2003], [Meyer et al. 2002] und [Postolache 2004] setzen sich mit der Auflösung von definiten NPs in der englischen Sprache auseinander.

Da im Folgenden auf neuere Ansätze zur AR eingegangen wird, soll hier der Aufbau der gängigen Systeme anhand Abbildung 3 veranschaulicht werden.

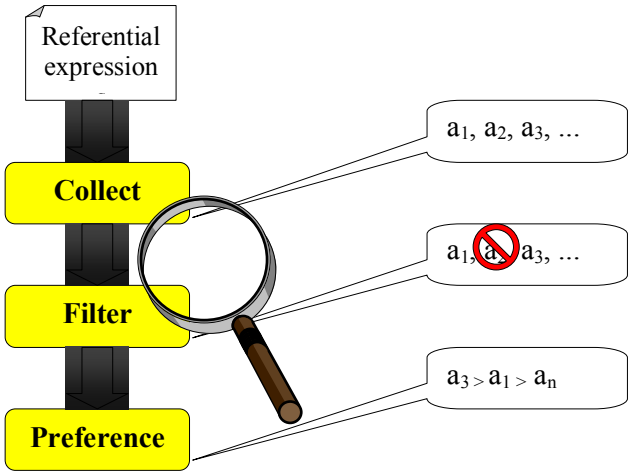

Abb. 3 Darstellung der Komponenten von AR-Systemen, basierend auf [Cristea 2001:26]

In einem ersten Schritt wird eine Liste mit möglichen Antezedenskandidaten zusammengestellt. Danach werden unpassende Kandidaten ausgefiltert und anschliessend wird der wahrscheinlichste Kandidat anhand von Präferenzen ermittelt.

[8] Ein POS-Tagger (Part-of-Speech-Tagger) ordnet jedem Wort die Wortart zu.
[9] Mit Parsern werden die Syntaxstrukturen eines Satzes ermittelt. Dabei wird die Eingabe anhand der Grammatik des Programms auf Korrektheit geprüft.
[10] Ein Korpus ist eine umfassende Textsammlung.
[11] Ontologien sind Wissensrepräsentationen. Es handelt sich hierbei um hierarchische Ordnungen von Begriffen.
[12] Mit Methoden des Maschinenlernens wird versucht, die Akquisition von Wissen zu automatisieren. Ein Sy stem kann aus Beispielen lernen und Verallgemeinerungen aufstellen, d.h. es werden Gesetzmässigkeiten in den Trainingsdaten erkannt. Nach der Lernphase kann das System auch unbekannte Daten beurteilen.
[13] Wahrscheinlichkeitsbasierte Ansätze

Wie bereits erwähnt, wurden auch in der Evaluation der AR-Systeme bedeutende Fortschritte erzielt. In Abbildung 4 sind die üblichen Messwerte aufgeführt.

[14]

$$\text{Präzision}(P) = \frac{\text{Anzahl der korrekt aufgelösten Anaphern}}{\text{Anzahl der aufgelösten Anaphern}}$$

$$\text{Ausbeute}(R) = \frac{\text{Anzahl der korrekt aufgelösten Anaphern}}{\text{Anzahl der aufzulösenden Anaphern}}$$

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

Abb. 4 Messwerte für die Evaluation von AR-Systemen, basierend auf [Byron 2004a:42] und [Jurafsky et al. 2000:578]

Unter *Präzision* versteht man den Quotienten der Anzahl der korrekt aufgelösten Anaphern und der Anzahl der aufgelösten Anaphern. Als *Ausbeute* bezeichnet man den Quotienten der Anzahl der korrekt aufgelösten Anaphern und der Anzahl der aufzulösenden Anaphern. Dabei ist zu beachten, dass unter den aufzulösenden Anaphern nicht alle Anaphern in den Testdaten zusammengefasst sind, sondern diejenigen, die der Algorithmus auflösen können müsste[15].
Präzision und Ausbeute sind tendenziell antagonistische Werte. Wenn ein System nach hoher Präzision strebt, wird seine Ausbeute tiefer ausfallen, im umgekehrten Fall höher[16]. Die *Performanz* ist ein Wert, der sich aus der *Ausbeute* (*R*) und der *Präzision* (*P*) zusammensetzt. β ist die Balance zwischen *Präzision* und *Ausbeute* und hat idealerweise den Wert 1. Das bedeutet, dass *Präzision* und *Ausbeute* gleichmässig gewichtet werden. Wenn β grösser 1 ist, wird *Präzision* bevorzugt, wenn β kleiner 1 ist, die *Ausbeute*.
Gemäss [Mitkov 2001:257] sollte ausserdem eine *nicht-triviale* und eine *kritische Erfolgsrate*[17] als Evaluationswert eingeführt werden, damit auch Aussagen über die Auflösung von schwierigen Anaphern gemacht werden können. Die *nicht-triviale Erfolgsrate* beinhaltet Anaphern, die mehr als einen Antezedenskandidaten haben. Die *kritische Erfolgsrate* enthält Anaphern, die nach der Anwendung von Genus- und Numerusfiltern mehr als einen Antezedenskandidaten haben.

[14] Im Englischen spricht man von *Precision* (Präzision), *Recall* (Ausbeute) und *F-Measure* (Performanz). Statt *Precision* wird auch der Begriff *Accuracy* verwendet.
[15] In einem Pronomenresolutionssystem werden also nur Pronomen – und nicht etwa auch vollständige NP-Anaphern – zu den aufzulösenden Anaphern gezählt.
[16] In einem perfekten System wäre das natürlich nicht der Fall; ein perfektes System hätte eine hohe Präzision und eine hohe Ausbeute.
[17] „Non-trivial success rate“ und „critical success rate”

Erfolgsrate =	$\frac{s}{n}$	s = n =	korrekt aufgelöste Anaphern alle Anaphern
nicht-triviale Erfolgsrate =	$\frac{s-k}{n-k}$	k =	Anaphern mit nur einem Antezedenskandidaten
kritische Erfolgsrate =	$\frac{s-k-m}{n-k-m}$	m =	Anaphern, die aufgrund von Numerus und Genus aufgelöst werden

Abb. 5 *Nicht-triviale* und *kritische Erfolgsrate*, basierend auf [Mitkov 2001:257]

[Barbu et al. 2001:1] argumentieren für eine gemeinsame Evaluationsplattform, da Unterschiede in den Evaluationsdaten und der Vorverarbeitung keine geeignete Vergleichsbasis darstellen. [Byron 2001:570, 574] schlägt Evaluationsschemata vor, in denen das Korpus, die Art der Anaphern, Auslassungen und Einschränkungen beschrieben werden.

Nachdem nun auf linguistische und computerlinguistische Grundlagen zum Phänomen Anaphora eingegangen wurde, werden im folgenden Kapitel verschiedene Typen von Anaphern beschrieben.

3 Verschiedene Typen von Anaphern

Wie kann man Anaphern unterscheiden? Es gibt verschiedene Typen von Anaphern, die ihrer syntaktischen Realisation, ihrer Lage, ihres Referenzvermögens oder ihrer Erschliessbarkeit nach unterschieden werden können. Abbildung 6 zeigt eine Übersicht der Unterscheidungsarten. Nähere Erläuterungen folgen unter 3.1 und 3.2.

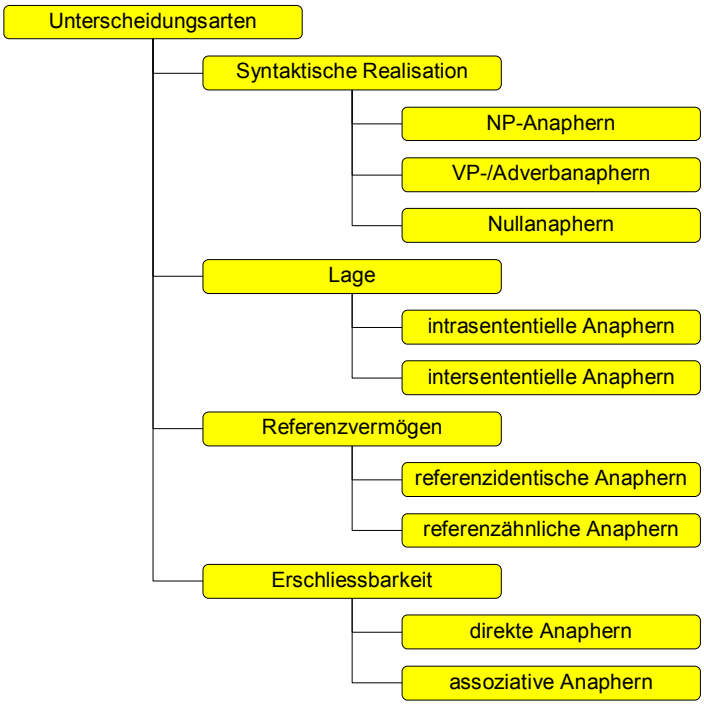

Abb. 6 Eigene Darstellung der Unterscheidungsarten von Anaphern

3.1 Unterscheidung nach syntaktischer Realisation

Anaphern können als NPs, VPs, Adverbien und Nullanaphern realisiert sein. NP-Anaphern können weiter in nicht-lexikalische Proformen und vollständige NPs unterteilt werden. Zu den nicht-lexikalischen Proformen werden in dieser Arbeit Pronomen und *One*-Anaphern gezählt, da sie lexikalisch gesehen leer sind und eine inhaltliche Spezifikation erst durch ihr Bezugswort erhalten. Unter vollständigen NPs sind Eigennamen, definite und indefinite NPs zusammengefasst. Im Gegensatz zu den nicht-lexikalischen Proformen haben vollständige NPs bereits eine Bedeutung unabhängig von ihrem Antezedens. Sie verweisen nicht nur auf bereits erwähnte Diskurseinheiten, sondern geben dem Rezipienten oft auch zusätzliche Information. Diese Arbeit konzentriert sich auf die Schwierigkeiten in der Auflösung von nicht-lexikalischen Proformen und vollständigen NPs, auf die in Kapitel 4 und 5 näher eingegangen wird.

Weit komplexer ist die Behandlung von Adverbanaphern (9a), VP-Anaphern (9b) und Nullanaphern (9c), und die Lösungsansätze sind, zumindest fürs Englische, eher rar[18].

(9) a *Will you walk with me to **the garden**? I've got to go down **there**...* [Mitkov 2002:12]

b *As the oil price **rose**, **so did** oil production.* [BNC: ABE 1388]

c *Ross carefully folded his trousers and ∅ climbed into bed.* [Peral et al. 2000:47]

*George was bought a box of **chocolates** but few ∅ were left by the end of the day.* [Mitkov et al. 2002c:13]

*I have never **been to Miami** but my father has ∅, and he says it was wonderful.* [Mitkov et al. 2002c:14]

Adverbanaphern können entweder lokal (9a) oder temporal (*then*) sein. Die Schwierigkeit besteht darin, dass sie, abgesehen vom lokalen und temporalen Aspekt, keine Gemeinsamkeit mit ihrem Antezedens aufweisen. Sie stehen also in keiner Synonymie-, Hyperonymie[19]-, Meronymie[20]-Beziehung etc. mit ihrem Antezedens. Dasselbe gilt für VP-Anaphern, die sich auf ein Verb oder eine Verbalphrase (VP) beziehen können. Eine weitere Schwierigkeit mit Adverben ist, dass sie auch deiktisch[21] verwendet werden können.

Nullanaphern (∅) sind unsichtbare Anaphern, die nicht durch ein Wort oder eine Phrase ausgedrückt werden, sondern sich als Lücke manifestieren, was ihre automatische Auflösung besonders erschwert. Pronomen, Hauptnomen einer NP oder VPs können ausgelassen werden, ohne das Verständnis der Aussage zu beeinträchtigen.

Angesichts der wenigen und noch nicht ganz ausgereiften Lösungsansätzen zu Adverbanaphern, VP-Anaphern und Nullanaphern liegt der Schwerpunkt dieser Arbeit auf der automatischen Behandlung von Proformen und vollständigen NPs.

3.2 Unterscheidung nach Lage, Referenzvermögen und Erschliessbarkeit

In Bezug auf die Lage kann man zwischen intrasententiellen und intersententiellen Anaphern unterscheiden. Von intrasententiellen Anaphern spricht man, wenn das Antezedens im selben

[18] In [Lappin 1990] und [Lappin 1996] sind Ansätze zur automatischen Verarbeitung von elliptischen VPs beschrieben, auf die hier nicht näher eingegangen wird.

[19] Überordnung; ein Hyperonym ist ein Oberbegriff eines anderen Begriffs, z.B. ist „Hund" der Oberbegriff für „Dogge".

[20] Teil-Ganzes-Beziehung; „Tischbein" ist ein Meronym von „Tisch"

[21] Bei deiktischen Verweisen bezieht sich ein Ausdruck nicht auf den Text, sondern auf die Sprechersituation, d.h. auf das Ich, Jetzt, und Hier des Sprechers. Wenn ein Sprecher beispielsweise eine Handgeste macht, während er den Satz „***There** was standing a house once*" äussert, wird mit dem *there* deiktisch auf die Stelle hingewiesen, wo einst ein Haus stand.

Satz wie die Anapher zu finden ist. Bei intersententiellen Anaphern liegt das Antezedens in einem vorangehenden Satz.

Ausserdem kann zwischen referenzidentischen und referenzähnlichen Anaphern unterschieden werden. Um referenzidentische Anaphern handelt es sich, wenn die Anapher und das Antezedens denselben Referenten in der realen Welt haben. Referenzähnliche Anaphern haben nur einen bedeutungsähnlichen Referenten. In Beispiel (10) sind zwei unterschiedliche Zahltage gemeint.

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

[Mitkov et al. 2002c:16]

Hinsichtlich ihrer Erschliessbarkeit kann man Anaphern in direkte und indirekte Anaphern einteilen. Indirekte Anaphern sind auch unter den Begriffen Bridging-Anaphern, assoziative Anaphern oder Inferrables bekannt. Fortan wird der Begriff assoziative Anaphora verwendet. Im Gegensatz zu direkten Anaphern können assoziative Anaphern nicht direkt aufgelöst werden, da ihr Referent nicht explizit im Text erwähnt wird. Der anaphorische Ausdruck kann aber über eine Brücke, das Antezedens, verstanden werden. Wie [Wind 2002:54] schreibt, ist das Antezedens nicht „referenzidentisch“ mit der assoziativen Anapher und „der Rezipient des Textes muss die Verbindung“ zwischen der assoziativen Anapher und dem Antezedens „erst herleiten“.

Es muss allerdings beachtet werden, dass in der Literatur keine Einigkeit über die Klassifikation von Anaphern herrscht. So werden beispielsweise assoziative Anaphern vorwiegend als anaphorisch eingestuft, es gibt aber auch Forscher wie [Bean et al. 1999:378] und [Gundel et al. 2005], die sie als nicht-anaphorisch ansehen. In dieser Arbeit werden assoziative Anaphern als anaphorisch betrachtet.

Nachdem nun die Unterscheidungsarten von Anaphern aufgezeigt wurden, werden im folgenden Kapitel die Schwierigkeiten mit den verschiedenen Typen in der AR näher beleuchtet.

4 Schwierigkeiten mit den verschiedenen Typen von Anaphern

[Cristea 2001:25] bringt die Schwierigkeiten in der AR auf den Punkt: „In order to do AR, one has to be able to do everything else. Once everything else is done, AR comes for free.“ Schwierigkeiten in der Vorverarbeitung haben einen grossen Einfluss auf die Erfolgsrate eines vollautomatischen AR-Systems. In der Vorverarbeitungsphase müssen morphologische Analysen durchgeführt, NPs extrahiert sowie Eigennamen und unbekannte Wörter erkannt werden. Ausserdem müssen pleonastische Pronomen[22] und nicht-anaphorische NPs vor dem eigentlichen AR-Prozess ausgefiltert werden, da ansonsten das System vergeblich versucht, ein Antezedens zu finden. Da diese Aufgaben nicht einwandfrei gelöst werden können, projizieren sich die Fehler auf die eigentliche Aufgabe der AR, was eine Reduktion der Erfolgsquote zur Folge hat.

In diesem Kapitel werden die Schwierigkeiten in der automatischen Behandlung von Proformen und vollständigen NPs erläutert.

4.1 Nicht-lexikalische Proformen und ihre automatische Behandlung

Unter dem Begriff nicht-lexikalische Proformen werden Pronomen und *One*-Anaphern zusammengefasst. Die automatische Auflösung von Pronomen wird durch verschiedene Faktoren erschwert.

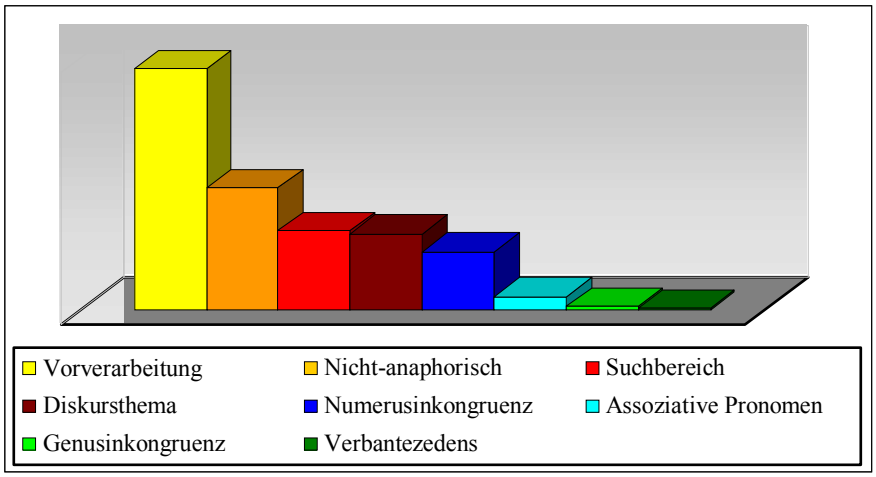

Abb. 7 Fehlerquellen in der Pronomenresolution, basierend auf [Evans 2002:55]

[22] Pleonastische Pronomen haben kein Antezedens im vorangehenden Text. Als Beispiel soll das Pronomen *it* in der Phrase *it rains* dienen, dessen Interpretation von keinem vorangehenden Diskurselement abhängt.

[Evans 2002:55] hat die Fehlerquellen in der Pronomenresolution analysiert und festgestellt, dass die meisten Fehler auf die Vorverarbeitung, die das Parsen, die Annotation und Textformatierung umfasst, zurückzuführen sind. Wie aus Abbildung 7 ersichtlich ist, bilden auch nicht-anaphorische Pronomen, die vor dem AR-Prozess ausgefiltert werden müssen, eine bedeutende Fehlerquelle. Dazu zählen kataphorisch, deiktisch und generisch[23] verwendete Pronomen sowie pleonastische Vorkommen. Eine weitere Reduktion der Erfolgsrate wird durch den beschränkten Suchbereich verursacht. In den meisten AR-Systemen wird im aktuellen Satz und den vorangehenden zwei Sätzen nach einem Antezedens gesucht. Es gibt aber auch Fernanaphern, deren Antezedenten weiter weg zu finden sind. Wie [Ariel 1988:70] aufzeigt, haben 81.3% der von ihr untersuchten Pronomen das Antezedens im gleichen oder vorangehenden Satz, während bei 18.7% der Pronomen das Antezedens innerhalb desselben Paragraphen oder sogar weiter entfernt zu liegen kommt.

Der Auflösungsprozess wird weiter durch Pronomen, die sich auf das Diskursthema beziehen, Numerus- und Genusinkongruenz sowie assoziative Pronomen erschwert. Die Problematik der Numerus- und Genusinkongruenz wird in einem eigenen Kapitel (Kapitel 5) behandelt, da nicht nur Pronomen, sondern auch andere linguistische Formen von Anaphern davon betroffen sind. Ausserdem haben die meisten Systeme Schwierigkeiten, wenn das Antezedens keine NP, sondern eine VP, ein (Teil-)Satz oder eine Abfolge von (Teil-)Sätzen ist. 16% der von [Gundel et al. 2005:351] untersuchten Personalpronomen der 3. Person hatten keine NP als Antezedens und im Trainingsset von [Evans 2001:13] war es ein Drittel der *it*-Pronomen.

Ein AR-System muss also nicht nur nicht-anaphorische Pronomen erkennen, sondern auch Antezedenten ermitteln, die nicht als NPs realisiert sind.

4.1.1 Erkennen und Ausfiltern von nicht-anaphorischen Proformen

Wie bereits erwähnt, müssen Pronomen nicht immer in einer anaphorischen Beziehung zu ihrem Bezugswort stehen. Verweise können auch kataphorischer, deiktischer und generischer Art sein und es gibt auch pleonastische Pronomen, die gar kein Antezedens haben.

4.1.1.1 Ausfiltern von kataphorischen, deiktischen und generischen Pronomen

Pronomen können auch auf ein Bezugswort vorverweisen, wie in Beispiel (7) gezeigt wurde. Ein Modul zur Erkennung von kataphorischen Pronomen kann daher zur Optimierung eines AR-Systems beitragen.

[23] Das Geschlecht oder die Gattung betreffend

[Evans 2001] versucht kataphorische Vorkommen von Pronomen zu klassifizieren, jedoch ohne Erfolg. Der Grund liegt darin, dass der Ansatz von [Evans 2001], der im Kapitel 4.1.1.2.2 beschrieben wird, zu sehr auf die Erkennung von pleonastischen Pronomen ausgerichtet ist. [Azzam et al. 1998:16] berücksichtigt Kataphern, jedoch nur in direkter Rede (11a) und in Konstruktionen mit Kopula[24] (11b)[25].

(11) a *„**I** caught Reggie when he was much younger counting his dad's trophies," **McNair** said.*
b ***He** is a **president**.*

[Azzam et al. 1998:16]

Für den Einbau eines Moduls zur Erkennung von Kataphern in ein AR-System sind für Kataphern typische Konstruktionsmuster notwendig, insbesondere wenn das System auf nicht manuell annotierten Daten basiert.
[Tanaka 1999:112] hat kataphorische Pronomen analysiert und festgestellt, dass sie oft in Zusammenhang mit direkter Rede (11a) und Adverbialphrasen (12) am Anfang des Satzes vorkommen.

(12) *In **his** news conference Wednesday, **President Carter** dismissed a suggestion raised Monday…*

[Tanaka 1999:112]

[Tanaka 1999:105-125] hat weitere subordinierende Konstruktionen mit kataphorischen Pronomen gefunden. Diese Konstruktionen sind allerdings weniger verbreitet und auch weniger typisch für Kataphora. So können kataphorische Pronomen in Konstruktionen mit *but* oder mit Rückversicherungsfragen vorkommen, wie Beispiele (13a) und (13b) zeigen.

(13) a ***His** beloved old white convertible was in terrible shape, but **my stepfather** refused to get rid of it.*
b *„Not the easiest place to sign a treaty on, is **it**?" he said, running his hand over **the rough surface of the table**.*

[Tanaka 1999:121,124]

Aber es gibt auch zahlreiche Beispiele von anaphorischen Pronomen, die in solchen Konstruktionen zu finden sind.

---

[24] Kopula sind Konstrukte, in denen ein Prädikat dem Subjekt zugesprochen wird oder das Subjekt das Prädikat ist. In dem Beispiel (11b) ist *He* das Subjekt, *is* die Kopula und *a president* das Prädikat.

[25] Man kann darüber streiten, ob es sich in Beispiel (11) wirklich um Kataphern handelt, da zumal (11b) keine Koreferenzbeziehung ist, und sich das Pronomen *he* wahrscheinlich auf eine vorangehende Einheit im Text beziehen würde.

Die Schwierigkeit, Kataphernresolution in den Prozess der AR zu integrieren, besteht hauptsächlich darin, dass die Grenze zwischen Anaphora und Kataphora nicht immer eindeutig ist und kataphorische Konstruktionen von komplexer Art sein können.

Nicht-anaphorisch ist auch die deiktische Verwendung von Pronomen. Obwohl oft in der gesprochenen Sprache deiktisch verwiesen wird, sind deiktische Verweise auch in Texten möglich, beispielsweise in Zusammenhang mit Abbildungen in Lehrbüchern (Beispiel (14)). Es stellt sich allerdings die Frage, wie solche Pronomen automatisch gefunden werden können.

(14) (Assuming that there is only one male figure in the picture)
*Now, look at the picture, what is **he** doing?*

[Tanaka 1999:42]

Nicht-anaphorisch ist auch der generische Gebrauch von Pronomen in Sprichwörtern und Redewendungen. Diese Pronomen können anhand von Listen ausgefiltert werden. Die Generierung von Listen kann aber sehr zeitaufwändig sein.

(15) a ***He** that plants thorns must never expect to gather roses.*
b ***He** who dares wins.*

[Mitkov et al. 2002c:10]

4.1.1.2 Ausfiltern von pleonastischen Pronomen

Neben kataphorisch, deiktisch und generisch gebrauchten Pronomen sind auch pleonastische Pronomen nicht-anaphorisch. Sie müssen dennoch in der AR erkannt werden, damit nicht vergeblich nach einem Antezedens gesucht wird. Pleonastische Pronomen kommen in Passivkonstruktionen mit modalen Adjektiven (16a) und kognitiven Verben (16b), in Ausdrücken von Witterungsverhältnissen (16c), Zeitausdrücken (16d), Distanzausdrücken (16e), Idiomen (16f), Spaltsätzen (16g) und Extrapositionen (16h) vor.

(16) a *It is dangerous, It is important, It is necessary…*
b *It is believed that, It appears that…*
c *It is raining, It is sunny…*
d *It is five o'clock, It is late, What day is today? …*
e *How far is it to Wolverhampton? It is a long way …*
f *At least we've made it, Call it quits, How's it going?*
g *It was Mr. Edgar who recruited Prudence Adair.* [Mitkov et al. 2002c:9, 10]
h *I just think it's so damn weird we're here.* [Gundel et al. 2005:359, 360]

In der Literatur herrscht allerdings Uneinigkeit bezüglich der Klassifikation von pleonastischen Pronomen vor. Laut [Denber 1998:10] lässt sich darüber streiten, ob pleonastische Pronomen einen Referenten haben (in Beispiel (16c,d) „das Wetter“ und „die Zeit“). [Byron 1999:54] argumentiert, dass es sich um ein referentielles Pronomen handelt, wenn das Pronomen durch *what* ersetzt und dabei eine Antwort gefunden werden kann. Für den Satz „*It's three p.m.*“ ist die Frage „*What's three p.m.*?“ und die Antwort „*The current time is three p.m.*“ [Byron 1999:54] sieht also *it*-Pronomen in Zeitausdrücken als referentiell an. Ein Antezedens hat das Pronomen dennoch nicht, der Referent kann aber aus der Zeitangabe inferiert werden. So gesehen steht das Pronomen in einer assoziativen Beziehung und ist somit anaphorisch.

Auch [Gundel et al. 2005:360] zeigen Beispiele, in denen es nicht ganz klar ist, ob das Pronomen tatsächlich nicht-anaphorisch ist. In Beispiel (17) kann das Pronomen *it* als pleonastisch angesehen werden. Es kann aber auch argumentiert werden, dass es sich auf die Phrase *the first to be called* bezieht.

(17) *My heart beat fast, for I had thought that as the discoverer of the body I would be the first to be called; but to my surprise,* ***it*** *was Marcel.*
[Gundel et al. 2005:360]

Ein weiterer Zweifelsfall ist im Beispiel (18) präsentiert. Hier kann das Pronomen *it* auch auf den Grund, das hohe Alter der Eltern, verweisen.

(18) A: *So that's why you're interested in death?*
B: *Maybe* ***it****'s because my parents were old? When I was young?*
[Gundel et al. 2005:360]

Im Gegensatz zu [Mitkov 2002a] und [Denber 1998] klassifiziert [Sayed 2003:3] *it*-Pronomen in Beispielen wie *„it seems that..., it is important to note that...*“ nicht als pleonastisch. [Sayed 2003:3] argumentiert, dass sich das Pronomen auf Fakten bezieht, die nach dem *that* stehen, wie in Beispiel (19).

(19) *What is unlikely? It is unlikely that the algorithm resolve all the anaphora.*
[Sayed 2003:3]

Wie können pleonastische Pronomen aber automatisch erkannt werden? Die ausführlichsten Algorithmen zur Erkennung von pleonastischen Pronomen sind die von [Paice et al. 1987],

[Lappin et al. 1994], [Denber 1998], [Evans 2001] und [Mollá et al. 2003], auf die im Folgenden näher eingegangen wird.

4.1.1.2.1 Erkennung von pleonastischen Pronomen mit Suchmustern

[Paice et al. 1987] versuchen pleonastische Pronomen in bestimmten Konstruktionen in technischen Texten zu erkennen. Ihr Algorithmus arbeitet mit Einschränkungen während des Mustervergleichs. So wird *it* als nicht-anaphorisch angesehen, wenn es in bestimmten Konstruktionen vorkommt, wie beispielsweise der Folge *it... to*. An den Text zwischen dem *it* und dem Konstruktionsende (hier: *to*) werden bestimmte Anforderungen gestellt. Der Text darf eine bestimmte Anzahl von Wörtern nicht überschreiten und Satzzeichen müssen beachtet werden.

Gemäss [Paice et al. 1987:120] dürfen maximal 27 Wörter zwischen der linken und der rechten Abgrenzung der Konstruktion stehen. Auch Interpunktionszeichen werden als Wörter gezählt. Es wird aber nicht über bestimmte Interpunktionssymbole wie „; : ? !)“ hinaus gesucht. Bei einer öffnenden Klammer wird der Text bis zur schliessenden Klammer ausgelassen. Ausserdem werden pleonastische Pronomen nicht direkt von Präpositionen wie *beside*, *to* und *upon* gefolgt und Präpositionen direkt vor dem Pronomen deuten auf anaphorische Vorkommen hin, wie Beispiel (20) zeigt. Die Wörter *so* und *for* werden aus dieser Präpositionsliste ausgeschlossen, da sie auch als Konjunktionen verwendet werden können.

(20) ***A stone** deserves to be examined after **it** has been scrubbed clean.*
[Paice et al. 1987:121]

Nun werden die *it*-Konstrukte, die auf pleonastische Vorkommen hindeuten und die entsprechenden Suchmuster näher betrachtet.

Das „*it..to*“-Konstrukt steht oft im Zusammenhang mit Aufgaben, wie im Beispiel (21). Dabei wird die Aufgabe mit der Phrase „*to limit the span of the header tanks*” ausgedrückt.

(21) ***it** is necessary **to** limit the span of the header tanks…*
[Paice et al. 1987:112]

Es kann also mit den Suchmustern in Abbildung 8 nach solchen Konstrukten gesucht werden. Dabei ist das VERB eine Form von „*to be, to seem, to appear,* etc.” Das STATUS- Wort bzw. die STATUS-Phrase beinhaltet Adjektive wie „*necessary, important, usual, difficult*” oder Phrasen wie „*of interest, of importance*”. Die Listen mit den entsprechenden Wörtern oder

Phrasen sind in [Paice et al. 1987: 130-132] zu finden. STATUS-Wörter, die sowohl in Zusammenhang mit nicht-anaphorischen als auch mit anaphorischen Vorkommen von *it* zu finden sind, wurden aus der Liste ausgeschlossen.

it VERB STATUS to TASK **VERB it STATUS to TASK... ?** **SUBJECT VERB it STATUS to TASK.**	**it VERB STATUS that STATEMENT** **it VERB CONTEXT that STATEMENT** **it VERB that**

Abb. 8 Suchmuster für *it..to*- und *it..that*-Konstrukte von [Paice et al. 1987:112,114,115]

Nach dem Konstrukt „*it..that*“ kann mit den Mustern in Abbildung 8 gesucht werden. Ein Beispiel für solche Konstrukte ist in Beispiel (22) gegeben.

(22) ***it*** *is in these cases* ***that*** *artificial respiration would offer the greatest chance of recovery*

[Paice et al. 1987:115]

STATUS-Wörter von „*it..that*“-Konstrukten sind Adjektive wie „*probable, essential, appropriate*“, Partizipien wie „*believed, shown*“ und indefinite NPs wie „*possibilities*“ [Paice et al. 1987:114]. Der CONTEXT ist üblicherweise eine Präpositionalphrase (PP), wie in Beispiel (22) gezeigt wurde. Die Bedingung, die an *it..that*-Konstrukte gestellt wird, ist, dass keine Präposition vor dem *that* vorkommt.

Weiter gibt es auch Konstrukte, die Unsicherheit ausdrücken, wie „*it..whether*“ und „*it..if*“ [Paice et al. 1987:116] und Sätze mit Fragewörter wie „*what, how, why, when, where*“. Ein Beispiel ist in (23) gegeben. In diesen Konstrukten sind STATUS-Wörter wie „*doubtful, not clear, unknown, wondered*“ enthalten.

(23) a ***it*** *is not known* ***whether*** *such an exchange was in fact performed…*
b ***it*** *might be questioned* ***what*** *he does learn*

[Paice et al. 1987:116]

Unsicherheit und Zweifel kann aber auch implizit in einem Satz enthalten sein, wie in (24). Diese pleonastischen Pronomen können aber kaum mit Suchmustern erkannt werden.

(24) ***it*** *is a matter for serious examination* ***how*** *so many under-age are able to see X films*

[Paice et al. 1987:116]

Weiter gibt es laut [Paice et al. 1987:117] auch „*it..which*“ und „*it..who*“-Konstrukte (25), die mit den Suchmustern in Abbildung 9 gefunden werden können. Die Bedingung ist, dass keine Präposition vor *which* steht.

(25) a ***it*** *was his Manchester period* ***which*** *produced the most important result…*
b ***it*** *is rather the teacher or parent* ***who*** *is moved by them*
[Paice et al. 1987:117,118]

it VERB CONTEXT which EVENT	**it VERB AGENT who ACTION**

Abb. 9 Suchmuster für *it..which* und *it..who*-Konstrukte von [Paice et al. 1987:117,118]

Zudem kann das Pronomen *it* auch in eingeschobenen Konstruktionen pleonastisch sein, wie in (26). Solche Konstruktionen sind durch das *it* und ein Komma abgegrenzt. Zwischen dem Pronomen und dem Komma dürfen aber nicht mehr als vier Wörter liegen.

(26) *Each of the 72 d has,* ***it*** *is now clear****,*** *a twin d'…*
[Paice et al. 1987:119]

Pleonastische Pronomen kommen auch in Zeitausdrücken und Zustandsbeschreibungen vor. Da aber solche Vorkommen in technischen Texten eher selten sind, wurden keine Regeln implementiert. Auch Konstrukte von *it* und Gerundium sind rar in technischen Texten und Vorkommen von *it*-Pronomen in Idiomen werden mit bestimmten Regeln behandelt.
[Paice et al. 1987:124] berichten von einer Erfolgsrate von 93.9%, die sie mit ihrem System über technischen Texten erreichten. Bei anderen Textgenres ist allerdings mit einer tieferen Erfolgsrate zu rechnen. Fehler sind auf die Unvollständigkeit der Listen, auf die Nicht-Implementierung von Regeln aufgrund mangelnder Beispiele in den Trainingsdaten und auf neue Idiome in den Testdaten zurückzuführen. Die Testergebnisse für die einzelnen Konstrukte können in [Paice et al. 1987:124-128] nachgelesen werden.

Im Ansatz von [Lappin et al. 1994] werden pleonastische Pronomen anhand der Suchmuster in Abbildung 10 erkannt. Die modalen Adjektive und die kognitiven Verben sind in Abbildung 11 aufgeführt. Natürlich werden auch Verneinungen sowie komparative und superlative Formen berücksichtigt.

It is **Modaladj** *that* **S**	**NP** *makes/finds it* **Modaladj** *(for* **NP***) to* **VP**
It is **Modaladj** *(for* **NP***) to* **VP**	*It is time to* **VP**
It is **Cogv-ed** *that* **S**	*It is thanks to* **NP** *that* **S**
It seems/appears/means/follows (that) **S**	

Abb. 10 Mögliche Konstruktionen, in denen das Pronomen *it* pleonastisch ist [Lappin et al. 1994:538]

Modale Adjektive	*advisable, certain, convenient, desirable, difficult, easy, economical, good, important, legal, likely, necessary, possible, sufficient, useful*
Kognitive Verben	*anticipate, assume, believe, expect, know, recommend, think*

Abb. 11 Modale Adjektive und kognitive Verben zur Erkennung von pleonastischen Pronomen [Lappin et al. 1994:538]

Das Problem in diesem Ansatz ist, dass die Listen wiederum nicht vollständig sind. [Denber 1998:22] schlägt daher vor, den Thesaurus WordNet[26] zu gebrauchen, um die Liste mit modalen Adjektiven zu ergänzen. Adjektive werden in WordNet mit attributiven Eigenschaften aufgeführt und können als modal betrachtet werden, wenn diese Eigenschaft einen Zustand, eine Bedingung, Qualität oder Quantität beschreibt.

[Denber 1998:22] sieht in Zusammenhang mit der Erkennung von pleonastischen Pronomen auch alle Adjektive mit der Endung *–ble* als modal an, obwohl es Adjektive wie „*variable*" und „*edible*" gibt, die keineswegs modal sind. Diese kommen allerdings nicht in Konstruktionen der Abbildung 10 vor.

Es ist allerdings anzumerken, dass dieser Vorschlag zur Erweiterung der Liste mit den modalen Adjektiven nicht implementiert wurde. Was die kognitiven Verben betrifft, werden diese in dem System von [Denber 1998:22] bereits mit dem verwendeten Tagger als solche markiert, und es ist daher nicht notwendig eine Liste aufzustellen.

[Denber 1998:20] versucht in seinem AR-Ansatz pleonastische Pronomen anhand der Suchmuster in Abbildung 12 zu erkennen.

Passivkonstruktionen	**Meteorologische Ausdrücke**	**Zeitausdrücke**
a *it* <*be*> modadj *that*	e *it* <*be*> ADV*PCP1	h *it* <*be*> ADV*NUM
b *it* <*be*> modadj [*for* N] *to* V	f *it* <*be*> ADJ	i *it* <*be*> PREP**time*
c N Vcog *it* modadj [*for* N] *to* V	g *it* <*be*> N	j *it* <*be*> ADV**(early|late)*
d *it* <*be*> modadj PRON		

Abb. 12 Suchmuster zur Erkennung von pleonastischen Pronomen von [Denber 1998:20-22]

[26] In WordNet sind Zusammenhänge zwischen Wörtern wie Synonymie, Antonymie, Hyponymie beschrieben. [Stand:22.8.2005] http://wordnet.princeton.edu/online/

In den Mustern kann das Verb *be* in einer flektierten Form vorkommen und der Kleene-Stern (*) bedeutet, dass der vorhergehende Ausdruck keinmal, einmal oder mehrmals vorkommen kann. Pleonastische Pronomen in Passivkonstruktionen wie „*it seems that..., it can be seen that..., it is clear that..., it seems he left...*" können laut [Denber 1998:21,22] mit den Suchmustern (a)-(d) der Abbildung 12 ausfindig gemacht werden. Pleonastische Pronomen in meteorologischen Ausdrücken werden mit den Mustern (e)-(g) gesucht.
PCP1 ist eine Liste mit 12 Verben, die in Zusammenhang mit Witterungsverhältnissen stehen, wie „*snowing, raining, drizzling*" [Denber 1998:21]. Mit dem Muster in (e) werden also Sätze wie „*It is hardly raining*" gefunden und ausgefiltert. Das Adjektiv ADJ aus Suchmuster (f) ist eines von 44 meteorologischen Adjektiven wie „*cloudy, sunny, windy*", die ebenfalls in einer Liste aufgeführt sind. In Suchmuster (g) ist das N eines der Nomen „*spring, summer, autumn, fall, winter*".
Nach pleonastischen Pronomen in Zeitausdrücken wird mit den Suchmustern (h)-(j) gesucht. Mit dem Suchmuster (h) werden Sätze wie „*It's three o'clock*" gefunden, da NUM ein Zeitausdruck ist. Das Suchmuster (i) matcht Sätze wie „*It's past time to go*" und das Suchmuster (j) Sätze wie „*It's late*" [Denber 1998:21].
Leider wurde das Modul zur Erkennung von pleonastischen Pronomen weder von [Lappin et al. 1994] noch von [Denber 1998] evaluiert.

Der Ansatz von [Mollá et al. 2003] ist dem Ansatz von [Lappin et al. 1994] sehr ähnlich, verwendet aber anstatt der SLOT-Grammatik die LINK-Grammatik. In der SLOT Grammatik hat jedes Wort eine bestimmte Anzahl von freien Stellen. Regeln legen fest, wie diese Leerstellen gefüllt werden dürfen. So verlangen Verben und Substantive beispielsweise bestimmte Präpositionen (*wait for, opinion about*). Die Grundidee der LINK-Grammatik besteht laut [Strüber 2001:2] darin, „jedem Wort im Lexikon eine Liste von Konnektoren zuzuordnen, die seine Bindungsmöglichkeiten beschreiben." Jede Verbindung (LINK) von zwei Wörtern hat einen Namen, der die Beziehung zwischen den Wörtern angibt. So trägt der LINK, der das Subjekt und das Verb verbindet, den Namen S, wie in Abbildung 13 gezeigt wird.

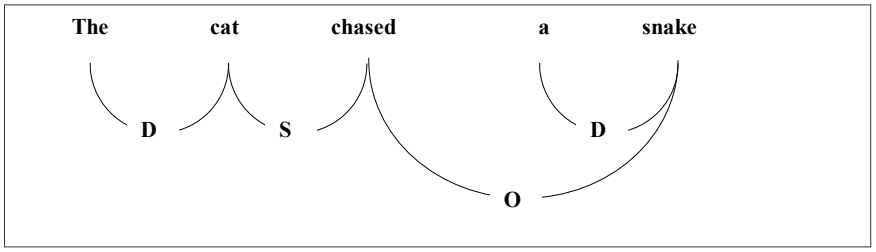

Abb. 13 Modellierung in der LINK-Grammatik, basierend auf [Strüber 2001:7]

Zur Erkennung von pleonastischen Pronomen verwenden [Mollá et al. 2003:70] die Konnektoren SF und SFI, welche die Füllsubjekte *it* und *there* mit Verben verknüpfen, wie in den Phrasen „*It is likely that...*“ und „*Is it likely that...?*“ [Mollá et al. 2003:70]. Der Algorithmus ignoriert absichtlich die Verbindungen SF und SFI, worauf pleonastische Pronomen gar nicht in den AR-Prozess miteinbezogen werden.

Problematisch mit den Ansätzen von [Paice et al. 1987], [Lappin et al. 1994] und [Denber 1998] ist, dass die Listen nicht vollständig sind. Wie [Sayed 2003:3,4] schreibt, muss das Pronomen *it* auch in Konstruktionen wie „*It is <pleasant | fine | okay | refreshing | tense | a tense situation>, It<is | would be | will be><fine | acceptable | a disaster> if <NP><VP>*” oder ähnlichen Ausdrücken als pleonastisch erkannt werden. Beispiel (27) dient zur Veranschaulichung.

(27) *Devising an ideal social and political system is next to impossible. It would be wonderful if we could do that.*

[Sayed 2003:4]

Zudem können einige Wörter manchmal auf den anaphorischen und manchmal auf den nicht-anaphorischen Gebrauch von Pronomen hindeuten. Ein weiterer Nachteil ist, dass die Suchmuster sehr einschränkend sind. Mit dem Suchmuster *It is* Cogv-*ed that* S (Abbildung 10) kann das pleonastische Pronomen *it* in der Konstruktion „*it was never thought by the committee that...*” [Evans 2001:6] nicht gefunden werden. Wenn die Suchmuster in einer weniger restriktiven Form angewendet werden, steigt allerdings die Anzahl von falsch klassifizierten pleonastischen Pronomen an.

Im Beispiel (28) kann das Pronomen von einem suchmusterbasierten AR-System fälschlicherweise als nicht-anaphorisch klassifiziert werden, da das Pronomen *it* gefolgt von *that* auf pleonastische Vorkommen hindeutet.

(28) *Studies of the ring chromosome that has XIST DNA but does not transcribe* ***it*** *show that its AR allele is…*

[Litrán et al. 2004:1]

Abhilfe für die eben genannten Probleme wird mit Techniken des MLs geschaffen.

4.1.1.2.2 Erkennung von pleonastischen Pronomen mit ML

[Evans 2001] hat einen Ansatz entwickelt, der nicht nur pleonastische Pronomen erkennt, sondern auch andere Vorkommen des Pronomens *it* zu klassifizieren versucht. [Evans 2001:3,4] unterscheidet *it*-Pronomen, die sich auf eine NP, einen (Teil-)Satz, eine VP oder ein Diskursthema beziehen sowie kataphorische, pleonastische und idiomatische Vorkommen. In einem ersten Schritt wird der Text mit einem FDG-Parser[27] analysiert. Das Programm liefert Informationen über die Wortklasse und das Lemma von Wörtern sowie Abhängigkeitsbeziehungen zwischen den Wörtern. Jedes Vorkommen von *it* wird als Vektor von 35 Merkmalen dargestellt, dem ein Wert zugeteilt wird. Anhand der Ähnlichkeit zu den Werten der Trainingsdaten wird das Pronomen klassifiziert. Die Pronomen des Trainingskorpus wurden manuell klassifiziert.

Die 35 Merkmale können gemäss [Evans 2001:11,12] in sechs Gruppen unterteilt werden. Gruppe 1 enthält Informationen über die Position des Pronomens. Es wird die Wortposition in einem Satz und die Satzposition in einem Paragraph erfasst. Die Annahme ist, dass nicht-anaphorische Pronomen Positionen am Anfang einnehmen, da sie sich auf keine vorangehenden Einheiten beziehen. Merkmale der 2. Gruppe beschreiben die Elemente des umliegenden Textes. So stehen beispielsweise pleonastische Pronomen selten direkt nach einer Präposition, werden aber oft von Adjektiven oder Komplementen gefolgt. In Gruppe 3 werden die Lemmata von umliegenden Verben und Adjektiven betrachtet. Mit dem Einbezug dieser Lemmata kann die Verwendung von Listen mit Wetterausdrücken und kognitiven Verben umgangen werden. Gruppe 4 enthält die Information über die POS-Tags der vier vorangehenden und vier nachfolgenden Wörter des Pronomens. Gruppe 5 enthält Konstruktionen, die auf pleonastische Pronomen hindeuten, wie beispielsweise Adjektive gefolgt von NPs. Merkmale der Gruppe 6 beschreiben schliesslich die Nähe der Pronomen zu Komplementen, zu Verben in der *-ing*-Form und zu Präpositionen.

[27] Functional Dependency Grammar

Wie [Evans 2001:15] schreibt, kann ihr System mit 67.94% Präzision und 89.19% Ausbeute Vorkommen von Pronomen, die sich auf eine NP beziehen, ermitteln. Pleonastische Pronomen können mit 73.38% Präzision und 69.25% Ausbeute erkannt werden. Das System erzielt allerdings schlechte Resultate in der Klassifizierung von Pronomen, die keine NP als Antezedens haben. Die Ursache liegt darin, dass die Merkmale zu sehr auf die Erkennung von pleonastischen Pronomen ausgerichtet sind. Ausserdem sind die Trainingsdaten unzureichend.
[Litrán et al. 2004] zeigen, dass Parsen nicht unabdingbar für die Erkennung von pleonastischen Pronomen ist. Ihr Ansatz, der auf dem Ansatz von [Evans 2001] basiert, macht zwar keinen Gebrauch von Parsing-Merkmalen, erzielt aber gleichwertige Ergebnisse. Die von [Litrán et al. 2004:3] verwendeten Merkmale sind in Abbildung 14 aufgeführt.

Merkmale	Beschreibung	Merkmale	Beschreibung
i	Line number	xiv	Lemma of previous verb
ii	Position in the line	xv	Lemma of next verb
iii-vi	POS tag of previous 4 words	xvi	Adjective + NP sequence
vii-x	POS tag of next 4 words	xviii	Number of previous prep.
xi	Distance to next compl.	xix	Number of following compl.
xii	Distance to next gerund	xx	Number of following adj.
xiii	Distance to next prep.	xxi	Number of previous NP

Abb. 14 Merkmale von [Litrán et al. 2004:3] zur Erkennung von pleonastischen Pronomen

Zusammenfassend kann man sagen, dass es zwei verschiedene Arten gibt pleonastische Pronomen zu erkennen. Suchmusterbasierte Ansätze sind insofern problematisch, als die Suchmuster zu einschränkend sein können, Listen nicht vollständig sind und bestimmte Wörter sowohl auf anaphorische als auch auf nicht-anaphorische Pronomen hindeuten.
Die neueren Ansätze mit Techniken des MLs scheinen viel versprechend zu sein. Ihr Vorteil liegt darin, dass Suchmuster nicht manuell generiert werden müssen und dass die Merkmale genreunabhängig sind, sofern ein genreübergreifender Testkorpus ausgewählt wird.

4.1.2 Erkennen von anaphorischen Proformen

Im Gegensatz zu nicht-anaphorischen Proformen haben anaphorische Proformen einen Bezug zu einer vorangehenden textuellen Einheit, die im AR-Prozess ermittelt werden muss. Abgesehen von den Schwierigkeiten mit den Diskrepanzen in den Kongruenzmerkmalen, die im Kapitel 5 beleuchtet werden, ist auch die Auflösung von *One*-Anaphern, Ereignis-Anaphern und assoziativen Pronomen problematisch. Eine besondere Behandlung erfordert ausserdem der Gebrauch von Pronomen in direkter Rede.

4.1.2.1 Automatische Auflösung von *One*-Anaphern

Obwohl es nur wenige Ansätze zur Auflösung von *One*-Anaphern gibt, ist ihre Bedeutung für die AR nicht zu unterschätzen. Wie [Dale et al. 2005:1] schreiben, liegt das Wort *one* schliesslich auf der Liste der häufigsten Wörter der englischen Sprache auf Rang 110.

Die Auflösung von *One*-Anaphern birgt einige Schwierigkeiten. Einerseits muss die anaphorische Verwendung von *one* (29c) vom nicht-spezifischen Pronomen (29a) und dem numerischen Gebrauch (29b) unterschieden werden. Andrerseits kann mit *one* eine Einheit aus einem Set ausgewählt werden (29c) und die *One*-Anapher muss daher mit ihrem Antezedens nicht im Numerus übereinstimmen. So kann mit der *One*-Anapher eine neue Diskurseinheit eingeführt werden, die allerdings anhand eines bereits bekannten Diskurselements interpretiert wird.

(29) a ***One*** *shouldn't pay more than twenty thousand dollars for an Acura.*
b *John has two Acuras, but I only have* ***one****.*
c *I saw no less than 6* ***Acura Integras*** *today. Now I want* ***one****.*
[Jurafsky et al. 2000:676]

Zudem kann sich die Anapher *one* sowohl auf ein Hauptwort der NP als auch auf die ganze NP beziehen, wie Beispiel (30) verdeutlicht.

(30) a *I don't think I'll have a sweet* ***pretzel****, just a plain* ***one****.*
b *I'll have* ***a sweet pretzel****. Mary takes* ***one*** *too.*
[Mitkov 2002:11]

One-Anaphern gehören zu den referenzähnlichen Anaphern, da das Antezedens und die Anapher nicht denselben Referenten haben, sondern nur einen ähnlichen.

4.1.2.1.1 Auflösung von One-Anaphern mit Ähnlichkeitskriterien und semantischen Hinweisen

[Ferrández et al. 1998:125] gehen *One*-Anaphern mit Ähnlichkeitskriterien an. Als Antezedens wird der Kandidat gewählt, der in Bezug auf die Syntaxstruktur der Anapher am ähnlichsten ist. In Beispiel (31) wird *T-shirt* gegenüber *Wendy* und *Sue* als Antezedens bevorzugt, weil es eine ähnliche Struktur aufweist, nämlich einen Artikel gefolgt von einem Adjektiv und dann ein Nomen.

(31) *Wendy didn't give either boy a green tie-dyed **T-shirt**, but she gave Sue a blue **one**.*

[Ferrández et al. 1998:125]

In manchen Fällen können semantische Hinweise dazu genutzt werden *One*-Anaphern automatisch aufzulösen. Im Beispiel (32) kommen in der automatischen Auflösung sowohl die Kleidungsstücke als auch die Preisschilder als Antezedenten der *One*-Anaphern in Frage. Semantische Hinweise für die automatische Auflösung sind die Farben und Stile, die die Kleider als Attribute haben. Antezedenskandidaten können also anhand ihrer Attribute bewertet werden. Es muss laut [Sayed 2003:3] daher geprüft werden, ob das Adjektiv *blue* auch ein Attributwert für *color* sein kann.

(32) *There were dresses of several different colors and styles. They were all pretty, labelled with price tags. Sally chose a blue one. Mary chose a skimpy one.*

[Sayed 2003:3]

4.1.2.1.2 Auflösung von One-Anaphern mit ML

Auch *One*-Anaphern können mit ML-Methoden angegangen werden. Einen solchen Ansatz bieten [Dale et al. 2005]. In einem ersten Schritt lernt das System die Vorkommen des Wortes *one* zu klassifizieren und in einem zweiten Schritt den anaphorischen Verwendungen ein Antezedens zuzuordnen.

[Dale et al. 2005] teilen die Vorkommen von *one* in numerische, partitive, anaphorische, generische, idiomatische und unklassifizierbare Verwendungen ein. Beispiel (33) dient zur Veranschaulichung des partitiven Gebrauchs. Mit *one* wird hierbei ein Teil aus einem nachfolgenden Set ausgewählt. *One* wird dabei von einer *of*-PP mit einem (Pro-)Nomen im Plural gefolgt.

(33) *A special exhibition of books for children forms **one** of the centrepieces of the 41st annual Frankfurt Book Fair.*

[Dale et al. 2005:2]

Unter dem generischen Gebrauch verstehen [Dale et al. 2005:2] das unbestimmte Pronomen *one* (siehe 29a). Als idiomatisch werden *one*-Verwendungen klassifiziert, die in idiomatischen Mustern zu finden sind, wie Beispiel (34) verdeutlicht.

(34) *It would be perfect to have a loved* ***one*** *accompany me.*

[Dale et al. 2005:2]

Die Vorkommen von *one* werden manuell in die entsprechenden Gruppen eingeteilt. Jedes *one* im Trainingsset wird als Vektor mit sieben Merkmalswerten dargestellt. Die sieben Merkmale werden kurz, basierend auf [Dale et al. 2005:3], erläutert. Die Merkmale können beispielsweise als Werte bestimmte POS-Tags, *Yes* oder *No*, *True* oder *False* haben.

Das erste Merkmal ist das POS-Tag. Das zweite Merkmal prüft, ob *one* von einer *of*-PP mit einem Hauptwort im Plural gefolgt wird. Ein drittes Merkmal ermittelt, ob sich *one* in der Subjektsposition befindet. Die Merkmale 4 und 5 überprüfen, ob nach *one*-Vorkommen in Subjektspositionen ein kognitives Verb wie *think*, *judge, doubt* oder ein Modalverb wie *must* steht. Merkmal 6 gibt die Position von *one* innerhalb der NP an. Dabei gibt es vier Möglichkeiten. Die NP kann nur aus *one* bestehen oder aber *one* befindet sich am Anfang, am Schluss oder in der Mitte der NP. Das siebte Merkmal ist das POS-Tag des vorangehenden Wortes.

Der numerische Gebrauch kann anhand der POS-Tags erkannt werden, da nur numerische Vorkommen als Adjektive markiert sind. Zur Erkennung des partitiven Gebrauchs verwendet man den syntaktischen Kontext und somit Merkmal 2. Unterscheidungen zwischen der generischen, idiomatischen und anaphorischen Verwendung von *one* sind komplexerer Art und es wird daher die Kombination der Merkmale berücksichtigt.

Das System kann die *one*-Vorkommen laut [Dale et al. 2005:3] mit einer Performanz von 88.9% klassifizieren. *One*-Anaphern werden mit einer Performanz von 73.7% erkannt. Natürlich beeinflussen Fehler im Taggen oder Parsen das Ergebnis.

Das Finden von Antezedenten für *One*-Anaphern basiert auf einem Trainingsset, das sich aus positiven sowie negativen Beispielen zusammensetzt. Positive Beispiele beinhalten korrekte Anapher-Antezedenspaare. Negative Beispiele werden generiert, indem die Anapher mit allen NPs, die zwischen ihr und dem korrekten Antezedens liegen, gepaart wird. Die Antezedenten der *One*-Anaphern wurden im Trainingsset ebenfalls manuell bestimmt. Es ist noch anzumerken, dass es auch *One*-Anaphern ohne explizites Antezedens gibt und deshalb eine entsprechende Markierung notwendig ist.

Das System lernt ebenfalls anhand von Merkmalen mit bestimmten Werten das korrekte Antezedens einer *One*-Anapher zu ermitteln. Das erste Merkmal ist das POS-Tag des Hauptwortes eines NP-Kandidaten. Das zweite Merkmal prüft, ob sich die *One*-Anapher und der Antezedenskandidat in einer PP befinden. Das dritte Merkmal untersucht, ob der Antezedenskandidat eine Subjektsposition einnimmt. Anhand des vierten Merkmals wird ermittelt, ob der Antezedenskandidat die zur Anapher nächstgelegene NP ist und Merkmal 5 schaut, ob sich der Kandidat in einem Relativsatz befindet.

Nach der Lernphase wird der Algorithmus über einem Testset angewendet. Die erreichte Performanz für das ganze System, d.h. die Identifikation und die Auflösung von *One*-Anaphern liegt bei 45,7%. Dies scheint viel versprechend zu sein, da es sich um den ersten Ansatz seiner Art handelt. Die Fehler aus dem Identifikationsalgorithmus werden in den Auflösungsprozess vererbt. Deshalb ist es das Ziel, die Klassifikation weiter zu verbessern, indem entweder weitere Merkmale berücksichtigt werden oder eine Nachbesserungs-Strategie als Abhilfe eingebaut wird.

4.1.2.2 Automatische Behandlung von Ereignis-Anaphern

Unter Ereignis-Anaphern sind in dieser Arbeit Anaphern zusammengefasst, die sich auf ein Ereignis, eine Situation, eine Aktivität, ein Faktum, eine Tatsache, oder eine vergleichbare Einheit beziehen. Ereignis-Anaphern haben einen (Teil-)Satz, eine Abfolge von (Teil-)Sätzen oder eine VP als Antezedens. In Beispiel (35) bezieht sich das Pronomen *it* auf das Ereignis der Mondlandung.

(35) *In 1969 America landed a man on the moon.* ***It*** *was NASA's shining moment.*
[Denber 1998:10]

Die automatische Auflösung von Ereignis-Anaphern ist jedoch problematisch, da es oft nicht ganz klar ist, auf was sich das Pronomen bezieht. In (36) kann sich das Demonstrativpronomen *this* auf die ganze Situation, die in (i), (ii) und (iii) beschrieben ist, beziehen oder nur auf (iii).

(36) (i) *John hated snakes*. (ii) *His neighbour had kept snakes,* (iii) *and he had been bitten by a snake once.* (iv) ***This*** *was unfortunate.*
[Schilder 1999:283]

Gemäss [Gundel et al. 2005:356] wird häufiger mit Demonstrativpronomen als mit Personalpronomen auf solche abstrakten Einheiten verwiesen. [Byron 2004a:10] bestätigt diese Aussage. In ihren Daten verweisen nur 39% der Personalpronomen, aber 66% der Demonstrativpronomen auf abstrakte Einheiten.

[Byron 2004a] hat sich der Auflösung von Pronomen in einem beschränkten Diskursmodell gewidmet. Das Besondere an dem Ansatz ist, dass auch Demonstrativpronomen aufgelöst werden können. Das Diskursmodell von [Byron 2004a:63] beinhaltet Diskurseinheiten und

abstrakte Einheiten, die den Inhalt von Konstituenten wie (Teil-)Sätzen und VPs repräsentieren. Dies soll an Beispiel (37) verdeutlicht werden.

(37) a *Each fall, penguins migrate to Fiji.*
b ***That**'s where they wait out the winter.*
c ***That**'s when it's too cold even for them.*
d ***That**'s why I'm going there next month.*
e ***It** happens just before the eggs hatch.*

[Byron 2002:80]

In (37b) und (37c) bezieht sich das Pronomen *that* auf die NPs *Fiji* beziehungsweise *fall* aus Satz (37a). In (37d) verweist *that* als Faktum auf den ganzen Satz in (37a), also auf die Wanderung der Pinguine nach Fiji im Herbst. In (37e) bezieht sich *it* auf das Ereignis der Migration der Pinguine. Um das Pronomen in (37d) und (37e) aufzulösen, müssen für (37a) die abstrakten Einheiten „Faktum" und „Ereignis" als Referenten bereitstehen.
Abstrakte Einheiten werden nach bestimmten Regeln generiert. Sätze, Teilsätze, untergeordnete Teilsätze, Infinitivphrasen und Gerundphrasen können abstrakte Einheiten bezeichnen, auf die ein Pronomen verweisen kann. Sprechakte sowie syntaktische und semantische Einschränkungskriterien werden verwendet, um dem Pronomen die zutreffende abstrakte Einheit zuzuordnen. So kann sich beispielsweise das Demonstrativpronomen *that* auf die Proposition des Satzes beziehen, falls es sich um eine Aufforderung (38a), nicht aber eine WH-Frage (38b) handelt.

(38) a A: *Send an engine to Elmira.*
B: ***That** will take three hours.*
b A: Why did an engine go to Elmira?
B: ****That** will take three hours.*

[Byron 2004:69]

Der Numerus gilt als syntaktische Einschränkung, während Verbbedeutungen semantische Einschränkungen sind. Das Demonstrativpronomen in dem Satz „***That** happened last night*" [Byron 2004a:76] muss sich aufgrund des Verbs *happen* auf ein Ereignis beziehen. Anhand der Abbildung 15 soll der Auflösungsprozess eines Demonstrativpronomens aufgezeigt werden.

Engine 1 goes to Avon to get the oranges.
(TELL (MOVE:theme x :dest y

: reason (LOAD : theme w)))
(the x (refers-to x ENG1))
(the y (refers-to y AVON))
(the w (refers-to w ORANGES1))

That *takes two hours.*

(TAKE-TIME: theme x:cost y)

Abb. 15 Auflösung eines Demonstrativpronomens von [Byron 2002:85]

Dabei schränkt das Prädikat TAKE-TIME seinen Inhalt auf ein Ereignis ein und MOVE ist ein Ereignis. Je nach Prädikat kann sich das Pronomen auf einen Teilsatz, einen Satz, eine NP etc. beziehen. Für eine detaillierte Beschreibung sollte [Byron 2004a:62-121] konsultiert werden.

Das System von [Byron 2004a] ist der erste implementierte Ansatz seiner Art, weshalb sich keine Vergleichsmöglichkeiten anbieten. Von 93 *that*-Pronomen konnten 62 aufgelöst werden. Das Diskursmodell ist aber beschränkt und es scheint unwahrscheinlich, dass grosse Textmengen mit diesem Ansatz aufgelöst werden können. Die Erkenntnisse von [Byron 2004a] liessen sich aber möglicherweise in anderen Ansätzen weiterverwenden.

4.1.2.3 Automatische Behandlung von assoziativen Pronomen

Unter assoziativen Pronomen werden Pronomen verstanden, deren Referent nicht direkt im Text erwähnt wird, aber implizit abgeleitet werden kann. Dabei dient das Antezedens als Interpretationshilfe. In Beispiel (39a) kann aus der Erwähnung der Kinder auf den impliziten Referenten, die Mutter, geschlossen werden. In (39b) verweist das ***it*** auf ein Dokument, das durch das Tippen entstanden sein muss.

(39) a ***Those kids*** *are just – And* ***she****'s pregnant with another one.* [Gundel et al. 2005:353]

b *When you have finished* ***typing****, press save, and store* ***it*** *in a suitable location.* [Evans 2002:56]

Die Ansätze, die sich mit der Auflösung von assoziativen Anaphern beschäftigen, beschränken sich vorwiegend auf definite NPs. Dies liegt wohl daran, dass assoziative Pronomen eher selten vorkommen, und wie [Gundel et al. 2005:354] schreiben, eher in spontanen Gesprächen

anzutreffen sind. Das Finden eines Antezedens für assoziative Pronomen ist besonders schwierig, da Pronomen leer sind und keine zusätzliche Information beinhalten.

4.1.3 Pronomenresolutionssysteme

Nachdem Schwierigkeiten in der automatischen Pronomenauflösung aufgezeigt worden sind, sollen nun einige gegenwärtige Pronomenresolutionssysteme vorgestellt werden. Zurzeit herrschen zwei Ansatzarten zur Behandlung von Pronomen in der AR vor. Wie bereits erwähnt, sind Ansätze am verbreitetsten, die aufgrund von syntaktischen und semantischen Merkmalen das Antezedens aus einer Liste von möglichen Kandidaten auswählen. Es gibt aber auch erste Ansätze, die Techniken des MLs verwenden.

4.1.3.1 Pronomenresolutionssysteme mit syntaktischen und semantischen Merkmalen

4.1.3.1.1 Das System von [Lappin et al. 1994]

[Lappin et al. 1994] haben einen Algorithmus entwickelt, um Pronomen in der 3. Person Singular aufzulösen, die eine NP als Antezedens haben. Der Algorithmus namens RAP operiert auf syntaktischen Strukturen der SLOT Grammatik.

RAP besteht aus mehreren Komponenten. Mit dem syntaktischen und morphologischen Filter werden unpassende Antezedenskandidaten ausgeschlossen. Pleonastische Pronomen werden mit der in Kapitel 4.1.1.2.1 beschriebenen Prozedur erkannt und in der Weiterverarbeitung nicht berücksichtigt. Der Anbindungsalgorithmus dient zur Auflösung von Reflexivpronomen und reziproken Pronomen. Mit dem Gewichtungs-Mechanismus werden den übrigen Antezedenskandidaten Präferenzen zugeschrieben. Ausserdem enthält RAP auch eine Prozedur zur Erfassung des Gesamtwertes der anaphorisch verlinkten NPs und ein Entscheidungsverfahren für die Auswahl des Antezedens aus einer Liste von möglichen Kandidaten.

Der syntaktische Filter besteht aus Einschränkungen, die Koreferenz zwischen einem Pronomen *P* und einer NP *N* innerhalb eines Satzes ausschliessen. Die Restriktionsbedingungen und Beispiele sind in Abbildung 16 dargestellt.

Einschränkungen	**Beispiele**
P und *N* stimmen in den Kongruenzmerkmalen nicht überein	**The woman$_i$ said that he$_i$ is funny.*
P ist im Argumentbereich von *N*	**She$_i$ likes her$_i$.*
P ist im Adjunktbereich von *N*	** She$_i$ sat near her$_i$.*
P ist ein Argument des Kopfes *H*, *N* ist kein Pronomen und in *H* enthalten	**He$_i$ believes that the man$_i$ is amusing.*
P ist im NP-Bereich von *N*	** John$_i$'s portrait of him$_i$ is interesting.*
P ist ein Artikel des Nomens *Q* und *N* ist in *Q* enthalten	**His$_i$ portrait of John$_i$ is interesting.*

Abb. 16 Syntaktischen Einschränkungen zur Ausschliessung von Koreferenz, basierend auf [Lappin et al. 1994:538]

Mit dem Anbindungsalgorithmus können mögliche intrasententielle Antezedenten für reziproke und reflexive Pronomen gefunden werden. Eine NP *N* ist das Antezedens eines reflexiven oder reziproken Pronomens *R*, wenn neben der Übereinstimmung in den Kongruenzmerkmalen eine der folgenden Bedingungen erfüllt ist:

Bedingungen	**Beispiele**
R ist im Argumentbereich von *N* und *N* füllt eine höhere Lücke als *R* in der Hierarchie (Oberflächensubjekt>Tiefensubjekt>Objekt) aus.	*They$_i$ wanted to see themselves$_i$.*
R ist im Adjunktbereich von *N*	*He$_i$ worked by himself$_i$.*
R ist im NP-Bereich von *N*	*John likes Bill$_i$'s portrait of himself$_i$.*
Es gibt eine artikellose NP *Q* im Argument- oder Adjunktbereich von *N* und *R* ist ein Argument von *Q*, oder es ist ein Argument einer Präposition, die ein Adjunkt von *Q* ist	*They$_i$ told stories about themselves$_i$.*
R ist ein Determinator eines Nomens *Q* und (i) *Q* liegt in dem Argumentbereich von *N* und *N* füllt eine höhere Argumentlücke als *Q* aus, oder (ii) *Q* liegt in dem Adjunktbereich von *N*	*[John and Mary]$_i$ like each other$_i$'s portraits.*

Abb. 17 Bedingungen an Antezedenten von Reflexivpronomen und reziproken Pronomen [Lappin et al. 1994:538,539]

Der Gewichtungsmechanismus schreibt jeder NP aufgrund von bestimmten Faktoren einen Wert zu. Als Gewichtungsfaktoren gelten Satznähe, Subjekte, existentielle Konstruktionen[28], Objekte, Hauptnomen und nicht-adverbiale NPs. Diese Gewichtungsfaktoren haben einen bestimmten anfänglichen Wert. Der Anfangswert für Satznähe ist beispielsweise 100, der Anfangswert für Subjekte 80. Die Anfangswerte sind in [Lappin et al. 1994:541] aufgeführt und werden im Verlauf des Prozesses vermindert. Mit jedem neuen Satz wird der aktuelle Wert für jeden Faktor halbiert. Wenn ein Faktor einen Wert nahe 0 erreicht, wird er entfernt. Auch jede koreferentielle Kette erhält eine Gewichtung. Dieser Gewichtungswert ist die Summe der aktuellen Gewichtung von allen Faktoren, in deren Reichweite mindestens ein Element der koreferentiellen Kette liegt.

[28] Konstruktionen mit *there is/there are*

Der Auflösungsalgorithmus läuft gemäss [Lappin et al. 1994:542-544] wie folgt ab. Zuerst werden alle NPs des laufenden Satzes nach ihrem Typ (definite NP, indefinite NP, pleonastisches Pronomen und anderes Pronomen) klassifiziert. Weiter werden alle NPs des laufenden Satzes untersucht und es wird zwischen NPs unterschieden, die neue Diskurseinheiten einführen, NPs, die auf bereits eingeführte Diskurseinheiten verweisen und NPs, die nicht-anaphorisch sind. Die Gewichtungsfaktoren werden auf die ermittelten Diskursreferenten angewendet. Der syntaktische Filter und der Anbindungsalgorithmus für reflexive und reziproke Pronomen werden angewendet und anaphorische Pronomen ausfindig gemacht. Wie pleonastische Pronomen erkannt werden, wurde bereits im Kapitel 4.1.1.2.1 erläutert. Wenn mehrere Kandidaten als Antezedens eines reziproken oder reflexiven Pronomens in Frage kommen, wird der Kandidat mit der höchsten Gewichtung gewählt.

Jedem Antezedenskandidaten für ein Personalpronomen wird ebenfalls ein Gewichtungswert zugeteilt, dieser kann sich aber noch verändern. So wird beispielsweise Kataphora bestraft, Parallelismus in Bezug auf grammatikalische Rollen hingegen begünstigt. Ausserdem werden nur Kandidaten berücksichtigt, deren Gewichtungswert über einem bestimmten Schwellenwert liegt. In einem weiteren Schritt wird die Kongruenz in den Übereinstimmungsmerkmalen geprüft und der syntaktische und morphologische Filter angewandt. Wenn mehr als ein Kandidat übrig bleibt, wird der Kandidat mit dem höchsten Gewichtungswert ausgewählt. Wenn mehrere Kandidaten den gleichen Wert aufweisen, wird der zur Anapher nächstgelegene Kandidat ausgewählt. Falls kein Singularkandidat gefunden werden kann, aber dafür ein Pluralkandidat gefunden wird, wird dieser als Antezedens betrachtet. Natürlich gilt dies auch im umgekehrten Fall.

In der Evaluation von RAP sind [Lappin et al. 1994:550,551] auf einige Einschränkungen ihres Algorithmus gestossen. So wird beispielsweise beim Gewichtungsverfahren nicht zwischen Haupt- und Nebensätzen unterschieden, was zu Fehlern bei der Auswahl des korrekten Antezedens führen kann. Im Beispiel (40) bekommen die Nomen *indicator* und *controller* dieselbe Gewichtung, da sie beide in der Subjektsposition stehen. RAP wählt schliesslich fälschlicherweise *controller* als Antezedens aus, weil es dem Pronomen *it* näher steht.

(40) *This green indicator is lit when the controller is on. It shows that the DC power supply voltages are at the correct levels.*

[Lappin et al. 1994:550]

Weiter kann auch das korrekte NP-Antezedens aufgrund seiner syntaktischen Position eine tiefe Gewichtung erhalten. Die korrekte NP kann beispielsweise in einer anderen NP einge-

bettet sein oder sich in einer anderen syntaktischen, für die Gewichtung ungünstigen, Position befinden.

[Lappin et al. 1994:550] schlagen deshalb statistisch gemessene lexikalische Präferenzen vor, um das System zu verbessern. Es wird der statistische Ansatz namens RAPSTAT von [Dagan et al. 1990] verwendet, der aufgrund von Kollokationsmustern aus einem Korpus jedem möglichen Antezedenskandidaten einen statistischen Wert zuweist, der mit der Gewichtungsbewertung von RAP verglichen wird. In Beispiel (41) sind *money*, *collection* und *government* mögliche Antezedenskandidaten für das *it* am Ende des Satzes. Das System prüft, welches Verb-Objekt-Paar im Korpus öfter vorkommt. Das Paar *collect money* ist den Paaren *collect collection* und *collect government* in der Anzahl seiner Vorkommen weit überlegen, weshalb *money* als korrektes Antezedens gewählt wird.

(41) *They know full well that the companies held tax money aside for collection later on the basis that the government said it was going to collect* ***it****.*
[Dagan et al. 1990:1]

Die möglichen Kandidaten K1, K2, K3 ... sind nach dem von RAP zugewiesenen Gewichtungswert in einer absteigenden Reihenfolge geordnet. RAPSTAT untersucht nun die ersten zwei Kandidaten K1 und K2. Wenn die Differenz der Gewichtungswerte von K1 und K2 nicht einen bestimmten Schwellenwert übersteigt und der statistisch zugeordnete Wert von K2 den von K1 übersteigt, wird K1 als möglicher Kandidat verworfen, während K2 bevorzugt wird. Die Liste mit den möglichen Kandidaten wird auf diese Weise abgearbeitet und der von RAPSTAT ausgewählte Kandidat als Antezedens herausgegeben.

Die Evaluation hat allerdings gezeigt, dass die Einbindung von RAPSTAT in RAP innerhalb des Genres von technischen Handbüchern nur zu einer mässigen Verbesserung führt. Die Erfolgsrate von RAP liegt gemäss [Lappin et al. 1994:554] bei 86%, die von RAPSTAT bei 89%.

Empfehlenswert ist die Online-Demoversion[29] des Ansatzes von [Lappin et al. 1994], in der auch die Schwierigkeiten in der Pronomenresolution veranschaulicht werden, insbesondere wenn komplexere Beispiele eingegeben werden.

[29] [Stand:22.8.2005] http://www-appn.comp.nus.edu.sg/%7Erpnlpir/cgi-bin/JavaRAP/JavaRAPdemo.cgi

4.1.3.1.2 Das System von [Denber 1998]

Auch [Denber 1998] hat ein Pronomenresolutionssystem entwickelt. Es handelt sich dabei um ein Bildsuchsystem, an das natürlichsprachliche Abfragen gestellt werden können. Auch dieses System beinhaltet einen Filter für pleonastische Pronomen, der im Kapitel 4.1.1.2.1 beschrieben ist. Das System filtert weiter ungeeignete Antezedenskandidaten anhand von Numerus-, Genus- und Belebtheitsmerkmalen und syntaktischen Rollen aus. Die Schwierigkeiten mit den Kongruenzmerkmalen Numerus, Genus und Belebtheit werden im Kapitel 5 näher beleuchtet. Übereinstimmung in der syntaktischen Rolle kann ein nützliches Einschränkungskriterium bei der Auswahl von Antezedenten sein. Im Beispiel (42a) bezieht sich das erste Pronomen *it*, das in der Subjektsposition steht, auf *the sun*, das Subjekt des ersten Satzes. Das zweite Pronomen *it* ist ein Objekt und verweist auf das Objekt *a lake* des ersten Satzes. Es gibt aber auch Gegenbeispiele, wie (42b), in denen die Anapher und das Antezedens nicht in der syntaktischen Rolle übereinstimmen.

(42) a *The sun is setting over a lake.* ***It*** *is reflected in* ***it****.*
b *At this point, EMACS is waiting for a command.* ***It*** *should be entered in lower-case.*

[Denber 1998:18]

Der Suchbereich des AR-Systems umfasst zwei Sätze und dasjenige Nomen, das mit den Pronomen in den meisten Kongruenzmerkmalen übereinstimmt, wird als Antezedens gewählt. Im Fall eines Unentschieden wird das Nomen, das mit dem Pronomen in der syntaktischen Rolle übereinstimmt, als Gewinner betrachtet. Wenn das Antezedens dennoch nicht ermittelt werden kann oder wenn dem Nomen oder Pronomen keine syntaktische Kategorie in der Vorverarbeitung zugewiesen wurde, wird das erste Nomen im vorangehenden Satz als Antezedens angesehen.

4.1.3.1.3 Das System von [Mitkov 2002a]

[Mitkov 2002a][30] stellt einen wissensarmen, robusten Ansatz zur Pronomenauflösung vor. Die erste Version geht auf das Jahr 1996 zurück, das System wurde allerdings stets weiterentwickelt und mittlerweile liegt eine vollautomatische Version namens MARS vor. Der Ansatz von [Mitkov 2002a:145-176] vermeidet komplexe Analysen und operiert über einem von einem POS-Tagger und NP-Extraktor vorverarbeiteten Text. Nach dem Einsatz eines syntaktischen Filters und einer Genus- und Numerusprüfung werden den möglichen

[30] Online Demoversion: [Stand:22.8.2005] http://clg.wlv.ac.uk/MARS/index.php

Antezedenskandidaten anhand von Antezedensindikatoren bestimmte Werte zugeordnet. Der Kandidat mit dem höchsten Gesamtwert wird als Antezedens gewählt. Im Fall eines Gleichstandes wird der Kandidat, der dem Pronomen am nächsten liegt, als Antezedens gewählt. MARS enthält zusätzlich bestimmte syntaktische Einschränkungen. So kann gemäss [Mitkov et al. 2002b:7] ein Pronomen nicht mit einem Ko-Argument[31] koreferieren. Zudem kann sich das Pronomen nicht auf eine nicht-pronominalen Konstituente, die es bestimmt und der es vorangeht, beziehen. Auch kann ein Pronomen nicht mit einer Konstituente, in der es enthalten ist, koreferent sein.

Die Indikatoren können den Antezedenskandidaten einen positiven oder negativen Wert zuordnen. Mit einem positiven Wert von +1 werden NPs in einem Abschnittstitel und NPs, die unmittelbar bestimmten Verben folgen, belohnt. Zu diesen indikativen Verben gehören „*analyse, assess, check, consider, cover, define, describe, develop, discuss, examine, explore, high-light, identify, illustrate, investigate, outline, present, report, review, show, study, summarise, survey, synthesise*" [Mitkov 2002a:146]. In der ursprünglichen Version erhalten erste NPs in einem Satz ebenfalls einen Wert von +1. In MARS hingegen wird gemäss [Mitkov et al. 2002b:6] NPs in Subjektposition ein Wert von +2 und NPs in Objektposition ein Wert von +1 zugewiesen. Indirekte Objekte erhalten keine Punkte und NPs, denen vom Parser keine grammatikalischen Funktionen zugeordnet werden, werden mit einem Wert von -1 bestraft. Syntaktischer Parallelismus wird in MARS mit einem Pluspunkt belohnt.

Die am häufigsten im Text vorkommenden NPs[32] und für das Genre typische Begriffe bekommen ebenfalls einen Wert von +1. Auch Pronomen sind als Antezedenten von Pronomen möglich und werden mit einem Wert von +1 honoriert. Die Zulassung von Pronomen als Antezedenten von Pronomen hat zwei Gründe. Einerseits wird mit pronominalen Verweisen die Aktualität von bestimmten Einheiten erhöht, d.h. je öfter auf eine bestimmte Einheit mit einem Pronomen verwiesen wird, desto mehr scheint diese Einheit Gesprächsgegenstand zu sein. Wie [Ariel 1996:24] schreibt, werden Pronomen gegenüber von vollständigen NPs bevorzugt, wenn auf das Diskursthema verwiesen werden soll. Andererseits kann das NP-Antezedens eines Pronomens ausserhalb des begrenzten Suchbereichs liegen und somit für das System nicht zugänglich sein. Pronomen-Antezedenten können daher als Trittsteine zwischen dem aktuellen Pronomen und seinem NP-Antezedens betrachtet werden. Natürlich kann dann mittels dieser Pronomen-Antezedenten auf das NP-Antezedens geschlossen werden.

---

[31] In dem Satz „*Peter gave him a parcel*" ist *gave* das Prädikat und *Peter*, *him* und *a parcel* sind (Ko)-Argumente. *Peter* und *him* können in diesem Beispiel also nicht-koreferentiell sein.

[32] frühere Versionen berücksichtigen die drei, MARS die zehn häufigsten NPs

Wenn eine NP einmal in dem Paragraph, in dem das Pronomen steht, wiederholt wird, wird ihr ebenfalls ein positiver Wert von +1 zugeordnet. Bei mehrmaliger Wiederholung erhält sie sogar einen Wert +2. Lexikalisch wiederholte Einheiten werden einerseits mit einfachen Suchen nach Zeichenfolgen erkannt. Andererseits wird aber auch nach NPs mit dem gleichen Kopf gesucht. Dabei werden sich ausschliessende Modifikatoren wie *first* und *second* berücksichtigt indem sie in einer Liste aufgeführt werden. Da der Ansatz von keiner Ontologie Gebrauch macht, kann allerdings nicht nach Synonymen oder Hyperonymen gesucht werden. In MARS werden aber nicht nur die Vorkommen einer NP gezählt, sondern auch die bisherigen Pronomen, die sich auf diese NP beziehen.
NPs, die im gleichen Kollokationsmuster stehen wie die Anapher, werden ebenfalls mit zwei Pluspunkten prämiert, wie Beispiel (43) aufzeigt.

(43) *Press **the key** down and turn the volume up… Press **it** again.*
[Mitkov 2002a:147]

Der Ansatz berücksichtigt die Kollokationsmuster „<noun phrase/pronoun, verb>, <verb, noun phrase/pronoun>“ und mit dem Verb to be „<noun phrase/pronoun, verb, adjective/past participle>“ [Mitkov 2002a:147]. Mit den Mustern „<(un)V, NP/pronoun>“ und „<NP/pronoun, (un)V>“ werden auch gegensätzliche Handlungen zugelassen, wie in der Phrase „*Loading **a cassette** or unloading **it***“ [Mitkov 2002a:147]. Ausserdem werden bestimmte Muster auch als Kollokationen betrachtet, wenn die Verbform ein Gerundium ist, wie im Beispiel „*When you plug in the power adapter, **the print head** moves to its protected position (you'll hear **it** moving)*“ [Mitkov 2002a:147].
Einen positiven Wert von +2 erhalten auch NPs in den Konstruktionen „...(You) V_1 NP... *con* (you) V_2 it (*con* (you) V_3 it)“ [Mitkov 2002a:147], wobei mit *con* eine Konjunktion wie *and*, *or*, *before*, *after*, *until* etc. gemeint ist. Wie aus dem Beispiel (44) ersichtlich ist, handelt es sich vorwiegend um imperative Konstruktionen und somit um eine genrespezifische Präferenz.

(44) *To print the paper, you can stand **the printer** up or lay **it** flat.*
[Mitkov et al. 2002c:148]

Weiter werden NPs in der NP_1-Position der Konstruktion „To V_1 NP_1, V_2 NP_2. (Sentence). To V_3 it, V_4 NP_4“ mit zwei Pluspunkten belohnt, da sie das wahrscheinliche Antezedens des Pronomens *it* sind. Als Beispiel dient (45).

(45) *To turn on **the video recorder**, press the red button. To programme **it,** press the 'Programme' key.*

[Mitkov et al. 2002c:148]

In MARS werden Kollokationsmuster anders bestimmt. Zuerst werden für jedes vorkommende Verb im Dokument die vorangehenden und nachfolgenden Hauptnomen (PHEAD und FHEAD) der NPs bestimmt und in einer Datei vermerkt. Abbildung 18 zeigt einen möglichen Dateiauszug.

VERB	eplace	PHEAD	you FHEAD it
VERB	replace	PHEAD	battery FHEAD cover
VERB	replace	PHEAD	printer FHEAD cartridge
VERB	replace	FHEAD	cartridge
VERB	replace	PHEAD	You FHEAD cartridge
VERB	replace	FHEAD	battery …

Abb. 18 Beispiel einer Liste mit den Verben und den vorangehenden und nachfolgenden Köpfen der NPs [Mitkov et al. 2002b:5]

Wenn nun das Pronomen *it* im Beispiel (46) aufgelöst werden soll, wird der NP *the battery* ein Wert von +2 zugewiesen, weil sie wie das Pronomen in Zusammenhang mit dem Verb *replace* in der Datei in der FHEAD-Position auftritt. Somit wird Information aus dem ganzen Dokument und nicht nur aus einem Paragraph verwendet.

(46) *Leave **the battery** inside the printer until you need to charge or replace **it**.*

[Mitkov et al. 2002b:5]

Minuswerte von -1 erhalten indefinite NPs und NPs in PPs. Es gibt ausserdem Indikatoren, die sowohl positive als auch negative Werte zuschreiben können. So werden beispielsweise dem Pronomen vorangehende NPs im selben Satz mit einem Wert von +2 honoriert, NPs im vorangehenden Satz erhalten +1 Punkte, die NPs im zweiten vorangehenden Satz werden weder belohnt noch bestraft, weiter entfernte NPs bekommen einen Wert von -1.

Der Nachteil dieser Indikatoren ist, dass sie teilweise sehr genrespezifisch sind. Ausserdem kann das korrekte Antezedens von einigen Indikatoren bestraft werden. Allerdings sollten alle Indikatoren zusammengenommen auf das korrekte Antezedens hindeuten.

Auf jeden Fall scheint sich der Ansatz zu bewähren, zumindest über technischen Texten. [Mitkov 2002a:151] berichtet von einer Erfolgsrate von 89.7% und einer kritischen Erfolgsrate von 82%. MARS wurde über einem Korpus von technischen Handbüchern evaluiert. Wie

[Mitkov et al. 2002b:10] schreiben, erreichte MARS eine Erfolgsrate von 59.35%, die mit dem Einsatz eines genetischen Algorithmus[33] auf 61.55% gesteigert werden konnte.

[Mitkov et al. 2002c] schlagen vor, mit einem zweisprachigen System die Auflösung von Pronomen zu verbessern[34]. Beispiel (47) soll dies veranschaulichen. In den meisten AR-Ansätzen werden NPs, welche die Rolle des Subjekts oder des direkten Objekts einnehmen, gegenüber NPs in der indirekten Objektsposition und innerhalb von PPs als Antezedens bevorzugt. Die meisten AR-Ansätze haben daher Schwierigkeiten das Pronomen *it* im Beispiel (47a) dem korrekten Antezedens *the videoplayer* zuzuordnen. Ein AR-System für das Französische hat allerdings aufgrund von Genusinformationen keine Probleme das äquivalente französische Beispiel (47b) aufzulösen.

(47) a *John removes the cassette from **the videoplayer** and disconnects **it**.*
b *Jean éjecte la cassette du **magnétoscope** et **le** débranche.*
[Mitkov et al. 2002c:201]

Ein bilinguales System zur Pronomenresolution kann sich bewähren, indem die eine Sprache zur Verbesserung der AR in der anderen Sprache genutzt wird. Das System von [Mitkov et al. 2002c:202] basiert auf alinierten Korpora[35] der englischen und französischen Sprache. Es handelt sich um die Übersetzung von technischen Texten vom Englischen ins Französische. Bei einer Übersetzung werden einige Pronomen der Ausgangssprache in der Zielsprache ausgelassen oder durch NPs und Demonstrativpronomen ersetzt. NPs der Ausgangssprache können in Pronomen der Zielsprache umgewandelt werden und es können sogar im Originaltext nicht vorhandene Pronomen in die Zielsprache eingeführt werden. Insbesondere die Übersetzung von einem Pronomen in eine NP kann für die AR sehr hilfreich sein.
Das englische Modul kann von der Genusinformation im Französischen profitieren. Im Französischen stimmt in den meisten Fällen das Pronomen in Bezug auf das Genus mit seinem Antezedens überein. Das französische Modul kann dem englischen auch helfen, wenn das englische Pronomen in eine französische NP übersetzt wurde. Wenn die französische NP eine englische Entsprechung hat, wird diese als Antezedens vorgeschlagen. Weiter kann auch die englische Entsprechung des im Französischen bevorzugten Antezedenskandidaten gewählt werden, wenn sich dieser von seinen Mitstreitern deutlich abhebt. Das bedeutet, dass eine

[33] Optimierungsalgorithmus basierend auf Heuristiken
[34] Mitkov verwendet in seinen Arbeiten oft technische Handbücher, die mehrsprachig zur Verfügung stehen
[35] Korpora, in denen die expliziten Entsprechungen zwischen der Ausgangssprache und der Übersetzung in eine Zielsprache markiert werden (z.B. durch Darstellung in Spalten).

Wertedifferenz von mindestens 4 Punkten zwischen dem höchstbewerteten Antezedenskandidaten und dem zweitbesten Kandidaten der französischen Sprache vorherrschen muss. Gemäss [Mitkov et al. 2002c:205] konnte die Auflösung von 38% der englischen Pronomen durch die Integration des französischen Moduls verbessert werden.

Auch das englische Modul kann zum Erfolg des französischen Moduls beitragen. Da der englische Algorithmus bisher besser ausgearbeitet ist als der französische, wird das Ergebnis des englischen Algorithmus höher gewichtet und im Fall einer Wertegleichheit wird der englische Kandidat bevorzugt. Das englische Modul kann bei der Auflösung von französischen Pronomen hilfreich sein, insbesondere wenn das französische Pronomen in eine englische NP übersetzt worden ist. Auch in Bezug auf kollektive Nomen[36] kann der englische Algorithmus zur Auflösung von französischen Pronomen beitragen. Während im Englischen auf kollektive Nomen mit den Pronomen *it* und *they* verwiesen werden kann, werden im Französischen die Singularpronomen *il* und *elle* verwendet. Wenn es im Französischen mehrere mögliche Antezedenskandidaten gibt, im Englischen Originaltext das Pronomen *they* auf ein Kollektivnomen verweist und keine weiteren Pluralkandidaten zur Verfügung stehen, kann die französische Entsprechung des englischen Kollektivnomens als Antezedens gewählt werden. Natürlich kann das englische Modul dem französischen helfen, wenn sich ein Antezedenskandidat, wie bereits oben erwähnt, besonders von seinen Mitstreitern absondert. Hier wird der Wert 3 als Differenz genommen.

In der Abbildung 19 sind die gegenseitigen Verbesserungen der Module aufgezeigt. Die Module wurden zuerst isoliert evaluiert und dann wurde das bilinguale System angewendet. Die Resultate zeigen die Erfolgsrate.

Total	***Vor gegenseitiger Verbesserung***	***Nach gegenseitiger Verbesserung***	***Verbesserung***
Englisch	72.25%	76.52%	4.27%
Französisch	58.21%	61.80%	3.59%

Abb. 19 Verbesserung der Erfolgsrate durch Anwendung eines bilingualen AR-Systems von [Mitkov 2002a:209]

4.1.3.1.4 Das System von [Siddharthan 2003]

[Siddharthan 2003] bietet einen Ansatz zur Auflösung von Pronomen an, der auf Suchmustern, Übereinstimmungsmerkmalen, Gewichtungsmechanismen und Syntaxfiltern basiert. Grammatikalische Funktionen wie Subjekte und Objekte werden mit Suchmustern gewonnen.

[36] Nomen, die Gruppierungen von Leuten, Tieren, Objekten etc. beinhalten (z.B. *a flock of sheep*)

Wie [Siddharthan 2003:8] schreibt, werden tendenziell Subjekte und in einem geringeren Ausmass auch direkte Objekte genreübergreifend als Antezedenten bevorzugt.

1	***obliq***	Prep NP_i	3	***dobj***	Verb NP_i
2	***subj***	NP_i [„[^Verb]+,” | „Prep NP”]* Verb	4	***iobj***	Verb [NP]+ NP_i

Abb. 20 Suchmuster für grammatikalische Funktionen von [Siddharthan 2003:8]

Mit diesen Suchmustern werden aber fälschlicherweise auch zeitliche Adjunkte in Sätzen wie „*The judge said Friday that...*“ [Siddharthan 2003:8] als Objekte gekennzeichnet. Dieser Fehler kann mit einer Regel, die Zeitausdrücke nur in Subjektspositionen als Antezedenten zulässt, bewältigt werden. Jedoch kann diese Regel zu einschränkend sein, wie aus dem Beispiel (48) hervorgeht. Der Algorithmus besitzt deshalb einen Mechanismus, mit dem solche Ausschlussregeln gelockert werden können, wenn kein Antezedens gefunden werden kann.

(48) *John prefers **Friday**. **It** is a convenient day.*

[Siddharthan 2003:8]

Antezedenskandidaten werden anhand von Kongruenzmerkmalen, die als Listen implementiert sind, ausgeschlossen. Abbildung 21 zeigt die verwendeten Merkmale und ihre möglichen Werte.

1	***number***	=	(s)ingular, (p)lural
2	***person***	=	(f)irst, (s)econd, (t)hird
3	***gender***	=	(m)ale, (f)emale, (n)euter
4	***animacy***	=	(a)nimate, (i)animate

Abb. 21 Kongruenzmerkmale von [Siddharthan 2003:9]

Wenn nicht genügend Information vorhanden ist, können die Merkmale auch unterspezifiziert sein. Mit Belebtheits- und Genusmerkmalen lassen sich Firmen und Tiere in diesem System elegant beschreiben. Eine Firma hat, wie [Siddharthan 2003:9] schreibt, die Merkmale „*gender* ={n} und *animacy* = {a}“, ein Tier die Merkmale „*gender* = {m, f, n}[37] und *animacy* = {a}.“ Das Pronomen *it* kann sich auf eine Einheit mit den Merkmalen *gender* = {n} und *animacy* = {a}, also eine Firma[38], ein Tier, oder aber auf eine Einheit mit dem Merkmal

[37] Im Englischen kann auf Tiere teilweise auch mit den Pronomen *he* und *she* verwiesen werden (vgl. (1))

[38] Firmen können als belebt angesehen werden, insbesondere wenn nicht das Unternehmen an sich, sondern die Mitarbeiter gemeint sind

animacy = {i} beziehen. Wie Genus- und Belebtheitsinformationen gewonnen werden können, wird in Kapitel 5 aufgezeigt. Weiter gibt es auch Einschränkungen in Bezug auf die direkte Rede. Ein Personalpronomen der 3. Person kann sich in der direkten Rede nicht auf den Sprecher beziehen.

Mit dem Ansatz von [Siddharthan 2003:10] können auch Relativpronomen aufgelöst werden. Die Kongruenzmerkmale für Relativpronomen sind in der Abbildung 22 aufgeführt.

	number	***gender***	***animacy***
who	{s,p}	{m,f,n}	{a}
which	{s,p}	{n}	{a,i}
that	{s,p}	{m,f,n}	{a,i}

Abb. 22 Kongruenzmerkmale für Relativpronomen, basierend auf [Siddharthan 2003:10]

Mit diesen Merkmalen ist gewährleistet, dass sich das Relativpronomen *who* nicht auf unbelebte Einheiten beziehen kann. Das Relativpronomen *which* kann auf Firmen, Tiere und unbelebte Einheiten verweisen, wobei mit dem Pronomen *that* eine Verweisung auf alle NPs möglich ist.

Die Antezedenskandidaten werden anhand der Kongruenzmerkmale bewertet. Für das Relativpronomen *who* werden menschliche Antezedenten mit den Merkmalen *animacy* = {a} und *gender* = {m,f} höher gewichtet, für *which* sind es Antezedenten, die Organisationen oder Tiere verkörpern und die Merkmale *animacy* = {a} und *gender* = {n} haben.

Neben den Kongruenzmerkmalen wird auch ein Syntax-Filter verwendet. Oft liegen nur PPs oder Appositionen[39] zwischen dem Antezedens und dem Relativpronomen, wie es auch der Fall in Beispiel (49) ist. Deshalb schliesst der syntaktische Filter alle möglichen Antezedenskandidaten aus, die durch eine andere Kategorie als eine PP oder eine Apposition, von dem Relativpronomen separiert werden. Wenn kein Antezedens gefunden werden kann, wird dieser restriktive Filter ausgeschaltet.

(49) *'The pace of life was slower in those days,'says* ***51-year-old Cathy Tinsall*** *from South London,* ***who*** *had five children, three of them boys.*

[Siddharthan 2003:10]

Die Evaluation des Ansatzes von [Siddharthan 2003:13] hat ergeben, dass die Akquisition von Genus- und Belebtheitsinformation aus dem umliegenden Text und WordNet die Resultate

[39] nähere Beschreibungen

genreunabhängig verbessern kann. Im Gegensatz hierzu führt die Anpassung von Gewichtungswerten nur zu genrespezifischen Verbesserungen.

4.1.3.2 Pronomenresolutionssysteme mit Techniken des ML

4.1.3.2.1 Das System von [Kehler et al. 2004]

[Kehler et al. 2004:1] stellen einen Ansatz zur automatischen Auflösung von Pronomen der 3. Person vor. Das Besondere an ihrem System ist, dass es sich selbst trainiert und dazu nicht einmal ein annotiertes Korpus benötigt. Wie [Kehler et al. 2004:2] schreiben, werden in einem ersten Schritt alle möglichen Antezedenskandidaten anhand von Numerus- und Genuseinschränkungen gefiltert. Wenn danach nur ein Kandidat übrig bleibt, wird die Verbindung zwischen dem Pronomen und dem Kandidaten als „wahr“ markiert. Andernfalls werden alle Verbindungen von möglichen Kandidaten-Pronomen-Paaren als „falsch“ herausgegeben.

Danach wird der Algorithmus „MaxEnt“ [Kehler et al. 2004:1] mit den im ersten Schritt herausgegebenen Daten trainiert. Der „MaxEnt“ schafft für jedes Pronomen in Verbindung mit jedem möglichen Antezedenskandidaten eine Trainingseinheit und gibt einen Wahrscheinlichkeitswert für die Verbindung aus. Der Wahrscheinlichkeitswert für jedes Pronomen-Kandidaten-Paar wird anhand von vierzig Merkmalen berechnet, die sich in die Gruppen Genus- und Numerusübereinstimmung, Distanz, grammatikalische Rolle und linguistische Form unterteilen lassen. Diese Merkmale sind liberaler als die Einschränkungen in Schritt 1 und so werden beispielsweise auch Kandidaten mit unbestimmten Genus- und Numerusinformation zugelassen.

Weiter wird der „MaxEnt“ auf jedes Pronomen-Kandidaten-Paar angewendet. Das Paar, das vom „MaxEnt” den höchsten Wahrscheinlichkeitswert erhält, wird als ‘wahr’ klassifiziert. Alle anderen Paar-Verbindungen werden als ‘falsch’ markiert.

Schliesslich wird der „MaxEnt”-Algorithmus mit diesen neuen Daten trainiert. Anschliessend werden die Schritte 3 und 4 solange wiederholt, bis die Trainingsdaten einen stabilen Zustand erreichen. Ein stabiler Zustand ist gemäss [Kehler et al. 2004:2] erreicht, wenn der „MaxEnt”-Algorithmus für ein Pronomen keinen anderen möglichen Antezedenskandidaten bevorzugt als in dem vorangehenden Durchlauf.

Zum Vergleich ihres selbst trainierenden Systems implementierten [Kehler et al. 2004:1] einen Algorithmus, der auf den gleichen Grundlagen beruht, aber auf einer vorannotierten Textsammlung trainiert wurde. Gemäss der Evaluation von [Kehler et al. 2004:3] wurde mit dem System, das sich anhand der annotierten Korpora selbst trainierte, nur leicht bessere Er-

gebnisse erzielt. Fehler des selbst trainierenden Ansatzes sind auf inkorrektes Parsen, Ausfilterung des korrekten Antezedens aufgrund der anfänglichen strengen Genus- und Numeruseinschränkungen und die Nicht-Behandlung von pleonastischen Pronomen zurückzuführen.

Nachdem die Lösungsansätze für die Auflösung von Pronomen aufgezeigt worden sind, widmet sich diese Arbeit nun den Problemen in der automatischen Behandlung von vollständigen NPs.

4.2 *Vollständige NPs und ihre automatische Behandlung*

Zu den vollständigen NPs gehören indefinite und definite NPs. Indefinite NPs haben entweder einen unbestimmten Artikel (*a car, some cars*) oder gar keinen (*cars*). Definite NPs beginnen mit einem bestimmten Artikel. Anaphorische definite NPs können weiter in Identitäts-Anaphern, assoziative Anaphern und *Other*-Anaphern unterteilt werden, die im Folgenden erläutert werden. Die Schwierigkeit bei der Auflösung von vollständigen NPs besteht einerseits darin, anaphorische NPs von nicht-anaphorischen zu unterscheiden und andererseits, das Antezedens der anaphorischen NPs zu finden.

4.2.1 Erkennen und Ausfiltern von nicht-anaphorischen NPs

Nicht-anaphorische NPs müssen von einem AR-System erkannt werden, damit das System nicht vergeblich nach einem Antezedens sucht. Insbesondere die Erkennung von nicht-anaphorischen definiten NPs spielt eine bedeutende Rolle in der AR, da zahlreiche definite NPs kein Bezugswort im Text haben. Im Korpus von [Bean et al. 1999:373] sind 63% der definiten NPs nicht-anaphorisch.

Es gibt verschiedene Arten von nicht-anaphorischen definiten NPs. Nicht-anaphorisch sind definite NPs, die deiktisch oder generisch gebraucht werden, oder aus dem Kontext- bzw. Weltwissen erschliessbar sind. Weiter können definite NPs auch neue Diskursinformationen einführen und daher nicht-anaphorisch sein. Wie können solche nicht-anaphorischen definiten NPs aber automatisch erkannt werden?

4.2.1.1 Erkennen und Ausfiltern von exophorisch verweisenden vollständigen NPs

Wie bereits erwähnt, müssen exophorisch verweisende vollständige NPs vor dem Auflösungsprozess ausgefiltert werden. Beispiel für den deiktischen Verweis ist die folgende

Äusserung eines Sprechers mit der entsprechenden Gestik: „***Those skirts** are lovely, aren't they*?" Beispiele für die Interpretation durch Weltwissen oder Kontextwissen sind die Ausdrücke *the sun* oder *the pope*. Hier ist jedoch anzumerken, dass diese definiten NPs in bestimmten Situationen auch anaphorisch gebraucht werden können.

Definite NPs, welche aufgrund von Kontextwissen oder Weltwissen verstanden werden, können mit Heuristiken angegangen werden. So verhalten sich diese definiten NPs oft wie Eigennamen („*the United States*"), haben Eigennamen als Modifikatoren („*the Iran-Iraq war*"), oder beziehen sich auf Zeitangaben wie „*the morning*" [Vieira et al. 2000:552]. [Vieira et al. 2000:553] halten die Zeitangaben in einer Liste fest, die Ausdrücke wie „*hour, time, morning, afternoon, night, day, week, month, period, quarter, year*" und die entsprechenden Pluralformen beinhaltet. Weiter gehören gemäss [Vieira et al. 2000:553] auch „*the moon*, *the sky*, *the pope*" und „*the weather*" zu den definiten NPs, die anhand von Weltwissen aufgelöst werden können.

Auch restriktive Modifikation[40] deutet auf nicht-anaphorische definite NPs hin, die anhand von Kontext- oder Weltwissen interpretiert werden können. Ein restriktiver Prämodifikator kann ein Nomen (50a), ein Adjektiv, eine Zahl (50b) oder ein Eigenname (50c) sein.

(50) a *A native of the area, he is back now after riding **the oil-field boom** to the top, then surviving the bust running an Oklahoma City convenience store.*
b ***the 1987 stock market crash*** [Vieira et al. 2000:556]
c ***the U.S. president*** [Bean et al. 1999:375]

Restriktive Postmodifikation wird oft durch PPs (51a), Relativsätze (51b) und Appositionen realisiert.

(51) a ***the president of the United States***
b ***the president who governs the U.S.***
[Bean et al. 1999:375]

Ausserdem können definite NPs generisch verwendet werden und müssen daher kein Bezugswort im Text aufweisen, wie Beispiel (52) zeigt.

(52) *No one knows precisely when **the wheel** was invented.*
[Mitkov et al. 2002c:11]

[40] Unter restriktiver Modifikation versteht man die Modifikation, die zur Identifizierung der Diskurseinheit unentbehrlich ist, wie beispielsweise restriktive Relativsätze und Präpositionalphrasen. Wenn die Modifikation nicht zur Identifizierung der Diskurseinheit beiträgt – die Diskurseinheit also anhand des Hauptnomens bestimmt werden kann – spricht man von nicht-restriktiver Modifikation.

Interessant sind die Beispiele in (53). Einerseits scheint die Konstruktion mit *of* eine definite NP zu bevorzugen. Andererseits scheint sich die Generizität in (53b) aus der eingebetteten indefiniten NP zu ergeben.

(53) a *the sum of 3 and 4*
b *the beauty of a sunset*

[Bosch 2001:2]

4.2.1.2 Erkennen und Ausfiltern von diskursneuen definiten NPs

Ebenfalls nicht-anaphorisch sind definite NPs, die neue Einheiten in den Diskurs einführen und nicht anhand von gemeinsamem Wissen zwischen den Diskursteilnehmern interpretiert werden können. [Vieira et al. 2000: 551] verwenden Heuristiken um diskursneue definite NPs zu erkennen. Es gibt Modifikatoren und Hauptnomen, welche diskursneue Information signalisieren. So deutet der Modifikator *first* in der Phrase „*the first person to sail to America*“ und das Nomen *fact* im Satz „*the fact that there is life on Earth*” [Vieira et al. 2000:551] auf neue Diskursinformation hin. Neue Information wird oft mit definiten NPs in Kopula-Konstruktionen (***the Prime Minister*** is Tony Blair) oder in Appositionen (*Glenn Cox, the president of Phillips Petroleum Co.* [Vieira et al. 2000:552]) eingeführt.
Definite NPs in der Subjektsposition der Verben „*to be, to seem* oder *to become*“ werden als diskursneu klassifiziert, wenn das Komplement des Verbs keine Adjektivphrase ist. Ebenfalls als diskursneu gelten definite NPs in der Objektposition des Verbs *to be*. Dies wird in (54) verdeutlicht.

(54) a *... **the man most likely to gain custody of all this** is a career politician named David Dinkins.*
b *What the investors object to most is **the effect they say the proposal would have**...*

[Vieira et al. 2000:557]

Modifikatoren und Nomen, die auf diskursneue Information hindeuten, können in Listen zusammengefasst werden. Die Liste der Modifikatoren von [Vieira et al. 2000:553] beinhaltet Wörter wie „*first, last, best, most, maximum, minimum, only*“ und Superlative, die ebenfalls nicht-anaphorische definite NPs signalisieren, wie (55) zeigt.

(55) *She jumps at **the slightest noise**.*

[Vieira et al. 2000:553]

In der Liste der Nomen sind „*fact*, *result*, *conclusion*, *idea*, *belief*, *saying*, *remark*" [Vieira et al. 2000:552] aufgeführt. [Vieira et al. 2000:552] identifizieren nicht-anaphorische definite NPs, indem sie den Kopf der NP oder ihre Modifikatoren mit Wörtern aus den Listen vergleichen.

Restriktive Modifikation ist laut [Vieira et al. 2000:554] das typischste Merkmal für diskursneue Einheiten. In (56) sind Beispielsätze für restriktive Modifikation gegeben.

(56) a *The place where he lives…*
b *The guy we met…*
c *The man writing the letter is my friend.*
d *The man to consult is Wilson.*

[Vieira et al. 2000:554,555]

Mit den Suchmustern in Abbildung 23 kann Postmodifikation erkannt werden, die auf diskursneue Information schliessen lässt.

[NP, the, _Premodifiers_,_Head_, [SBARQ |_] |_]; [NP, the, _Premodifiers_,_Head_, [SBAR |_] |_]; [NP, the, _Premodifiers_,_Head_, [S |_] |_]; [NP, the, _Premodifiers_,_Head_, [VP |_] |_];	[NP, the, _Premodifiers_,_Head_, [PP,_ |_] |_]; [NP, the, _Premodifiers_,_Head_, [WHPP,_ |_] |_]. [NP, [NP, the, _Premodifiers_,_Head_], [Clause]].

Abb. 23 Suchmuster zur Erkennung von Postmodifikation, die auf diskursneue Information hindeutet, von [Vieira et al. 2000:555]

Nicht-restriktive Postmodifikation, die nicht auf diskursneue Information hinzudeuten scheint (57), kann anhand von Kommata erkannt werden.

(57) *The substance, discovered almost by accident, is very important.*

[Vieira et al. 2000:555]

Eigennamen mit vorangehendem Artikel (58) werden bei ihrem ersten Vorkommen als nicht-anaphorisch klassifiziert. Bei weiteren Erwähnungen werden sie als anaphorisch angesehen.

(58) a *the Securities and Exchange Commission* [Vieira et al. 2000:557]
b *the H_2O molecule* [Bosch 2001:2]

4.2.1.3 Algorithmus zur Erkennung von nicht-anaphorischen NPs

Wie aufgezeigt wurde, gibt es zahlreiche nicht-anaphorische NPs, die vor dem Auflösungsprozess ausgefiltert werden müssen. Heuristiken zur Erkennung solcher NPs wurden soeben beschrieben. Wie aber kann man diese Heuristiken als Algorithmen formulieren?

[Bean et al. 1999] haben ein System zur automatischen Erkennung von nicht-anaphorischen definiten NPs entwickelt. Ihr Algorithmus generiert aus einem Trainingskorpus eine Liste von nicht-anaphorischen definiten NPs und NP-Mustern, die zur Ermittlung von nicht-anaphorischen definiten NPs in anderen Texten verwendet werden kann.

[Bean et al. 1999:375] verwenden Heuristiken, um nicht-anaphorische definite NPs von anaphorischen zu unterscheiden. Zuerst werden nicht-anaphorische definite NPs anhand von syntaktischen Hinweisen ermittelt. Wie bereits erwähnt, kann restriktive Prä- und Postmodifikation auf nicht-anaphorische definite NPs hindeuten. Definite NPs der Form „*the* <number><noun>“ wie in „*the 12 men*“ [Bean et al. 1999:375] und solche, deren Kopf im vorangehenden Text erscheint, werden als anaphorisch klassifiziert. In dem Korpus von [Bean et al. 1999:373] konnten 24% der nicht-anaphorischen definiten NPs mit syntaktischen Hinweisen erkannt werden.

Dann wird versucht, nicht-anaphorische definite NPs anderweitig zu ermitteln. Dazu werden die definiten NPs, die in dem ersten Satz eines Textes vorkommen und nicht mit syntaktischen Heuristiken klassifiziert werden können, in einer Liste 1 zusammengefasst und als nicht-anaphorisch eingestuft. Weiter werden aus der Liste 1 modifizierte Hauptnomen, die nicht-anaphorische definite NPs repräsentieren, wie beispielsweise „*the Salvadoran Government*“ [Bean et al. 1999:375] extrahiert. Diese Hauptnomen werden anhand von Mustern der Form „the <x+><noun1...nounN>“ gesucht, die anhand des Algorithmus in Abbildung 24 generiert werden. Die gefundenen definiten NPs werden in der Liste 2 aufgeführt.

1. Für jede NP, die aus mehr als zwei Wörtern besteht, schaffe ein Muster der Form „the <x+> headnoun.“ Beispiel: Bei der NP „the new Salvadoran government“ ist das Muster „the <x+> government“
2. Wende dieses Muster auf das Korpus an und zähle wie häufig es auf eine NP zutrifft.
3. Falls möglich, erweitere das Muster, indem das Wort zur Linken des Hauptnomens der NP ins Muster eingefügt wird. Beispiel: „the <x+> Salvadoran government“
4. Wende das ergänzte Muster auf das Korpus an, zähle wie häufig es auf eine NP zutrifft. Falls die Anzahl dieses Musters geringer ist als die des ursprünglichen Musters, stoppe hier und kehre zum ursprünglichen Muster zurück. Falls die Anzahl gleich der des ursprünglichen Musters ist, kehre zu Schritt 3 zurück. Mit diesem sich wiederholenden Prozess werden zusammengesetzte Hauptnomen erkannt.

Abb. 24 Algorithmus zur Generierung von Hauptnomen-Mustern zur Erkennung von nicht-anaphorischen definiten NPs von [Bean et al. 1999:376]

Weiter gibt es auch nicht-anaphorische NPs, welche nur in definiter Form auftreten und nie oder äusserst selten in indefiniten Konstruktionen zu finden sind. Um diese definiten NPs aus dem Korpus zu akquirieren, wurden zuerst alle Hauptnomen von NPs ermittelt und entsprechend ihren Vorkommen in definiten und indefiniten Konstruktionen geordnet. Aus den Hauptnomen, die mindestens 5 Mal im Trainingskorpus auftauchten und immer in definiten Konstruktionen zu finden waren, wie die NP „*the contrary*“ [Bean et al. 1999:376], wurde eine Liste 3 erzeugt.

Ein Korrektur-Algorithmus musste eingebaut werden, da es auch definite NPs gibt, die im ersten Satz anaphorisch sein können und so irrtümlich in die Liste 1 gelangen. Zudem gibt es auch definite NPs, die sowohl nicht-anaphorisch als auch anaphorisch sein können, wie die NP „*the guerillas*“ [Bean et al. 1999:376], die in einigen Texten ein Antezedens hat und in anderen nicht.

Der Korrektur-Algorithmus zur Beseitigung dieser Fehlerquellen basiert auf der Wahrscheinlichkeit, mit der eine NP in einer definiten Konstruktion vorkommt. Wenn dieser Wahrscheinlichkeitswert eine obere Schwellengrenze überschreitet, darf die NP als nicht-anaphorisch klassifiziert werden. Wenn der Wahrscheinlichkeitswert eine Untergrenze nicht erreicht, darf die NP nicht als nicht-anaphorisch eingestuft werden. Und wenn der Wert zwischen dem unteren und dem oberen Grenzwert liegt, kann die NP als „manchmal nicht-anaphorisch“ gekennzeichnet werden.

Abbildung 25 veranschaulicht den Algorithmus zur Erkennung von nicht-anaphorischen definiten NPs. Die Listen 1, 2 und 3 werden anhand eines Trainingskorpus generiert.

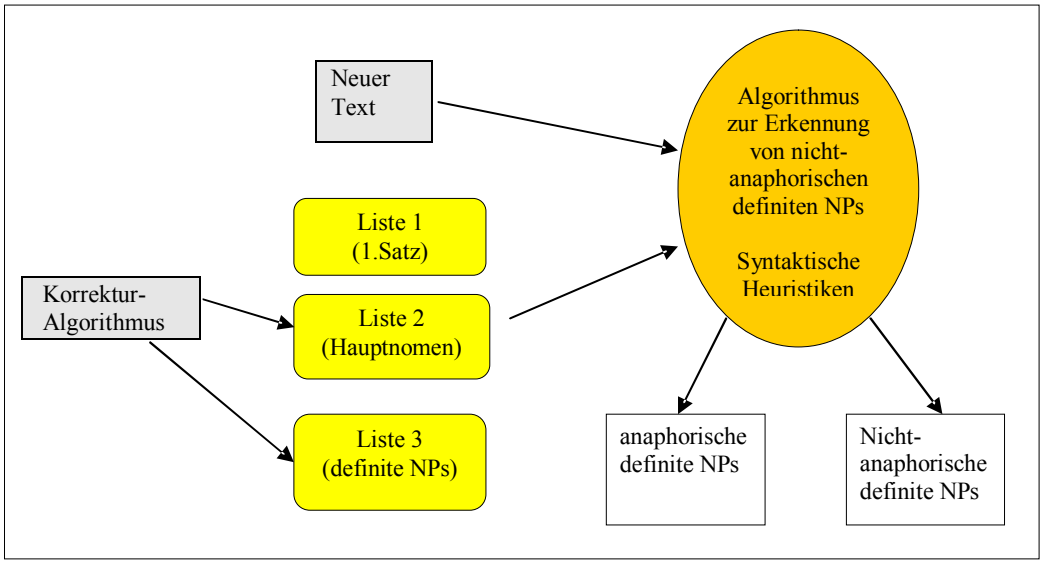

Abb. 25 Algorithmus zur Erkennung von nicht-anaphorischen definiten NPs, basierend auf [Bean et al. 1999:377]

Der Algorithmus prüft jede definite NP in einem Text. Zuerst werden die Heuristiken zur Erkennung von anaphorischen definiten NPs angewendet und die NP wird allenfalls als anaphorisch klassifiziert.

Falls die NP nicht als anaphorisch klassifiziert werden kann, wird nach ihr in den Listen 1, 2 und 3 gesucht. Wenn die NP in einer der Listen gefunden wird, wird sie als nicht-anaphorisch markiert. Andernfalls werden die Heuristiken zur Erkennung von nicht-anaphorischen definiten NPs angewendet und die NP wird allenfalls als nicht-anaphorisch klassifiziert. Falls die NP immer noch nicht klassifiziert werden kann, wird sie als anaphorisch betrachtet.

Mit ihrem Ansatz erreichten [Bean et al. 1999:378] eine Ausbeute von 77.7% beziehungsweise 79.1% und Präzisionswerte von 86.6% respektive 84.5%, je nach der Festsetzung der Schwellenwerte. Der Nachteil des Algorithmus ist, wie [Cardie et al. 2002:1] schreiben, dass die lexikalische Information, die zur Generierung von Listen mit nicht-anaphorischen NPs gebraucht wird, genrespezifisch ist.

4.2.2 Automatische Auflösung von anaphorischen NPs

Unter anaphorischen vollständigen NPs sind hier Identitätsanaphern, assoziative NP-Anaphern und *Other*-Anaphern zusammengefasst, auf die im Folgenden näher eingegangen wird.

4.2.2.1 Automatische Auflösung von Identitätsanaphern

Bei Identitätsanaphern handelt es sich um definite NPs, die sich auf die gleiche Diskurseinheit wie das Antezedens beziehen. Wie aus (59) ersichtlich ist, kann dabei das Hauptnomen nochmals aufgegriffen oder ein Synonym verwendet werden.

(59) a *Fred was discussing* ***an interesting book*** *in his class. I went to discuss* ***the book*** *with him afterwards.*
b *Fred was wearing* ***trousers****.* ***The pants*** *had a big patch on them.*
[Bunescu 2003:1]

Eine Identitätsanapher kann auch einen Namen als Antezedens haben. So kann mit der NP *the Prime Minister* auf den zuvor erwähnten *Tony Blair* verwiesen werden. Das System benötigt allerdings hierfür die Information, dass *Tony Blair* auch *Premierminister* ist. Diese Information kann aus einem Namenerkennungssystem gewonnen werden.

Ausserdem kann auch eine Hyperonymie- oder Hyponymiebeziehung zwischen der Anapher und dem Antezedens vorherrschen, beispielsweise wenn zuerst von einem Hund und an-

schliessend von einer Dogge (oder umgekehrt) die Rede ist. Hier ist anzumerken, dass die Grenze zwischen Identitätsanaphern und assoziativen Anaphern nicht eindeutig ist und Hyperonymie- oder Hyponymiebeziehungen auch zu den assoziativen Anaphern gezählt werden können, wie es [Poesio et al. 2002:1221] und [Vieira et al. 2000:558] tun. In dieser Arbeit werden sie als Identitätsanaphern aufgeführt, da der Rezipient im Gegensatz zu assoziativen Anaphern nur eine Einheit vor Augen hat.

Identitätsanaphern können ebenfalls mit Heuristiken angegangen werden. [Vieira et al. 2000:547] konzentrieren sich auf Identitätsanaphern mit dem gleichen Kopf. Ihr System sucht nach möglichen Antezedenten, deren Kopf mit dem der definiten NP übereinstimmt. Es muss dabei allerdings beachtet werden, dass es auch definite NPs ohne Hauptnomen gibt, wie „*the highest in the southern region*“ [Vieira et al. 2000:547]. Zudem müssen nicht nur die Köpfe, sondern auch die Modifikatoren übereinstimmen. So kann beispielsweise die NP „*the red car*“ nicht mit der NP „*the blue car*“ übereinstimmen, obwohl der Kopf identisch ist. Postmodifikatoren müssen hierbei ebenfalls beachtet werden. So stehen die Phrasen „*a chance to accomplish several objectives*“ und „*the chance to demonstrate an entrepreneur*“ [Vieira et al. 2000:551] nicht in einer anaphorischen Relation.
[Vieira et al. 2000:551] haben deshalb die folgenden Heuristiken eingeführt: Eine definite NP bezieht sich auf ein Antezedens, wenn die Modifikatoren der definiten NP eine Teilmenge der Modifikatoren des Antezedens sind. Ein nicht-modifiziertes Antezedens kann mit irgendeiner definiten NP übereinstimmen, wenn der Kopf übereinstimmt. Mit diesen Heuristiken können Anapher-Antezedenspaare, wie „*an old Victorian house ... the house*“ und „*a check... the lost check*“ [Vieira et al. 2000:551] gefunden werden.
Vorteilhaft am System von [Vieira et al. 2000:550] ist, dass auch eine NP ausserhalb des fixen Suchbereichs als Antezedens zugelassen werden kann, falls es sich um eine Identitätsanapher handelt.

Es gibt aber auch Identitätsanaphern, die in einer komplexeren Beziehung als Synonymie und Hyperonymie zu ihrem Antezedens stehen. Wie sollen solche Fälle von Identitätsanaphern aufgelöst werden?
Wenn mit einer definiten NP auf eine andere NP verwiesen wird und die beiden NPs nicht im Lemma übereinstimmen, kann dies zu Schwierigkeiten im Auflösungsprozess führen, wie in Beispiel (60) verdeutlicht wird.

(60) *Amenomphis the IVth's wife… the beautiful queen*

[Postolache 2004:8]

Wie [Postolache 2004:8] schreibt, können in solchen Fällen neben Ansätzen, die auf Ähnlichkeitskriterien beruhen, auch das natürliche Geschlecht sowie die Modifikatoren des Hauptnomens zur korrekten Auflösung beitragen. Die Königin ist einer Ehefrau ähnlicher als einem Pharao, vorausgesetzt dem System ist bekannt, dass Amenomphis ein Pharao ist. Aber auch wenn dies nicht bekannt ist, kann anhand eines Namenerkennungsprogramms zumindest festgestellt werden, dass es sich um eine männliche Person handelt. Die Königin ist abermals der Ehefrau ähnlicher als einem Mann. Es gibt bereits Ansätze, welche anhand von Ontologien Ähnlichkeitsbeziehungen messen, jedoch sind diese noch nicht fehlerfrei.

In manchen Fällen, kann das natürliche Geschlecht (≠ grammatikalisches Genus) in der AR von Bedeutung sein. So sind im Beispiel (60) die Königin sowie die Ehefrau weiblichen Geschlechts. Auch die Modifikatoren des Hauptnomens können bei der Auflösung hilfreich sein, wie Beispiel (61) zeigt. Das Adjektiv *beautiful* deutet auf das Nomen *beauties* hin.

(61) *the most beautiful women… those beauties*

[Postolache 2004:8]

NP-Anaphern können sich sowohl im Lemma als auch im Numerus von ihrem Bezugswort unterscheiden. Auf Inkongruenz in Bezug auf den Numerus wird im Kapital 5.1 näher eingegangen.

Wie bereits gezeigt wurde, können NPs in Appositionen und Kopula-Konstrukten nicht-anaphorisch sein, wie in Beispiel (54). Ebenfalls nicht-anaphorisch ist (62b). (62a) ist hingegen anaphorisch und es stellt sich die Frage, wie automatisch zwischen anaphorischen und nicht-anaphorischen Kopula-Konstrukten unterschieden werden soll.

(62) a *The young lady became a wife.*
b *Dick Cheney is not the president of the United States.*

[Postolache 2004:10,11]

Nachdem die Probleme und Schwierigkeiten mit Identitätsanaphern vorgestellt worden sind, wird nun die Thematik der assoziativen NP-Anaphern aufgegriffen.

4.2.2.2 Automatische Auflösung von assoziativen Anaphern

Von assoziativen NP-Anaphern spricht man, wenn sich die Anapher nicht auf die gleiche Diskurseinheit wie das Antezedens bezieht, aber über das Antezedens auf die Bedeutung der Anapher geschlossen werden kann. Im Beispiel (63) hat die definite NP *the front door* kein explizites Bezugswort im Text. Häuser haben aber üblicherweise Türen und der Referent wird somit indirekt über das Antezedens inferiert.

(63) *Jane has a new house.* ***The front door*** *is blue.*
[Wilson et al. 1998:173]

Manchmal ist auch spezifisches Wissen oder Weltwissen notwendig, um die Anapher zu identifizieren. In Beispiel (64a) muss bekannt sein, dass Athos, Porthos und Aramis Musketiere sind. In Satz (64b) kann man aus der Zeitangabe *8 p.m.* auf die Dunkelheit schliessen.

(64) a ***The musketeers*** *are attacked on the bridge.* ***Athos*** *is wounded, but* ***Porthos and Aramis*** *succeed to win.* [Postolache 2004:11]
b *I left at* ***8 p.m. The darkness*** *made me jumpy.* [Clark 1977:171]

Es kann aber auch sein, dass eine definite NP auf ein Verb verweist und das Hauptnomen der NP vom Verb abgeleitet werden kann, wie in den Sätzen in (65).

(65) a *I* ***ran*** *two miles the other day.* ***The run*** *did me good.* [Clark 1977:170]
b *John* ***tried to convince Mary of his love. The attempt*** *was not successful.* [Sayed 2003:4]

[Sayed 2003:5] schlägt vor, dass solche Aktionsnomen, die sich von Verben ableiten lassen, in einer Liste mit den dazugehörenden Verben aufgeführt werden (z.B. *demonstration* → *demonstrate*). Für die Verben dieser Liste wird in WordNet nach Synonymen gesucht. Wenn eine definite NP aufgelöst werden muss, deren Hauptnomen ein Aktionsnomen ist, sucht der Algorithmus nach einem VP-Antezedens, das entweder das zum Aktionsnomen dazugehörende Verb oder ein Synonym dessen ist.
Die assoziative Anapher muss allerdings nicht immer eine definite NP sein. Im Beispiel (66) ist es die indefinite NP *a page*, die mit dem Buch assoziiert werden kann.

(66) *I picked up that* ***book*** *I bought and* ***a page*** *fell out.*
[Prince 1992:305]

Auch NPs ohne Artikel können anaphorisch sein, insbesondere in technischen Gebrauchsanweisungen oder in Kochrezepten.

Es können verschiedene assoziative Beziehungen zwischen einer Anapher und ihrem Antezedens vorherrschen. So kann die Anapher wie in (66) ein Teil des Antezedens sein. Hier spricht man von Meronymie-Beziehungen. Es kann sich aber auch um eine Zugehörigkeitsbeziehung (64a), eine Teilmengenbeziehung (67a) oder eine Element-von-Beziehung (67b) handeln.

(67) a *They saw **many coloured butterflies**. **The pink ones** were their favourites.*
b *They took **all the weapons for the underwater hunting**... **the masque**... **the rifle**... **the ribbon paws**.*

[Postolache 2004:12]

Es gibt auch komplexere Beziehungen, wie beispielsweise die zwischen „*Ausschwitz*“ und „*the victims*“ [Bunescu 2003:1] oder Ursache-Folge-Beziehungen. Zudem kann es auch vorkommen, dass das Bezugswort nicht explizit im Text erwähnt wird, aber implizit durch das Diskursthema erschliessbar ist. Als Beispiel nennt [Poesio et al. 2002:1221] „*the industry*“ in einem Text, der von Ölfirmen handelt.
Weiter ist anzumerken, dass assoziative Anaphern auch mehrere Bezugswörter haben können und diese nicht immer eindeutig sind. In (68) kann sich die NP *the windows* sowohl auf das Haus als auch auf den Wagen beziehen, weshalb Kontextwissen nötig ist, um das korrekte Antezedens zu bestimmen.

(68) *I drove by our house in my car. **The windows** were dirty.*

[Dorrepaal 1990:4]

Wie nun gezeigt worden ist, können assoziative Beziehungen von komplexer Art sein. Wie können assoziative Anaphern also automatisch aufgelöst werden?
Es gibt in der gegenwärtigen Literatur verschiedene Ansatzmöglichkeiten, um assoziative NP-Anaphern anzugehen. Neben WordNet-Ansätzen gibt es Ansätze, die auf Suchmustern oder Ähnlichkeitskriterien beruhen, und auch Techniken des MLs werden angeboten.
Oft wird bei der Selektion des Antezedens nach Konflikten zwischen den semantischen Beschreibungen der Anapher und den möglichen Kandidaten gesucht. Konflikte herrschen vor, wenn die Köpfe der NPs nicht in einer assoziativen Beziehung zueinander stehen. Weiter treten Konflikte auf, wenn die Modifikatoren nicht kompatibel sind und in einigen Fällen ist

auch Weltwissen oder Kontextwissen erforderlich, um die Kompatibilität der Modifikatoren zu prüfen.

4.2.2.2.1 Auflösung von assoziativen Anaphern mit Heuristiken und WordNet

[Vieira et al. 2000:557] versuchen assoziative Anaphern mit Heuristiken anzugehen und machen dabei von WordNet Gebrauch. Sie konzentrieren sich auf assoziative Anaphern, die in lexikalischen Beziehungen wie Synonymie, Hyperonymie und Meronymie stehen (69a), und auf solche, deren Antezedens ein Eigenname (69b) oder ein modifizierendes Nomen ist (69c)[41].

(69) a ***the flat*... *the living room***
b ***Bach*... *the composer***
c *the company has been selling* ***discount packages*... *the discounts***.
[Vieira et al. 2000:558]

Das System von [Vieira et al. 2000:559] zeigt eine assoziative Beziehung zwischen zwei Nomen an, wenn sie in einer Synonymie-, Hyperonymie- oder Meronymie-Beziehung zueinander stehen oder wenn sie Schwestern sind, was bedeutet, dass sie beide Hyponyme des gleichen Hyperonyms sind. Wie [Vieira et al. 2000:559] gemerkt haben, erfordert die Suche nach assoziativen Relationen komplexe Suchvorgänge in der WordNet-Hierarchie und ist nicht immer erfolgsversprechend. Für die Auflösung von assoziativen Anaphern, die sich auf einen Namen beziehen (69b), spielt Namenserkennung eine bedeutende Rolle und Personennamen, Firmennamen, Ortsnamen, Städtenamen, Staatsnamen etc. müssen erkannt werden.

Der Algorithmus zur Eigennamenerkennung von [Vieira et al. 2000:560] verwendet WordNet, das wichtige Namen enthält, erkennt aber auch Eigennamen anhand von Abkürzungen wie *Mr., Co,* und *Inc*. Mit einem Backtracking-Mechanismus[42] können ausserdem Namen, die noch nicht klassifiziert werden konnten, später einer Kategorie zugeordnet werden. Wie [Vieira et al. 2000:560] schreibt, kann beispielsweise das erste Vorkommen des Wortes *Morishita* nicht klassifiziert werden. Erst wenn der Name *Mr. Morishita* nochmals aufgegriffen wird, kann erkannt werden, dass es sich um einen Familiennamen handelt.

Das Antezedens einer assoziativen Anapher kann auch ein Kompositum sein, wie in „*stock* ***market*** *crash*... ***the markets***" [Vieira et al. 2000:560]. Der Algorithmus zur Auswahl von Antezedenten sucht daher nicht nur nach Übereinstimmungen zwischen den Hauptnomen der

[41] (69b) und (69c) können auch als Beispiele für Idenitätsanaphern angesehen werden
[42] Wenn keine passende Lösung gefunden wird, kann nach Alternativen gesucht werden, indem der Beweisgang zum letzten Entscheidungspunkt zurückkehrt.

Anapher und des Antezedens, sondern lässt auch vorangehende NPs als Antezedenten zu, wenn ihr Prämodifikator mit dem Kopf der definiten NP übereinstimmt. Auch wenn der Prämodifikator einer definiten NP mit dem Prämodifikator einer vorangehenden NP übereinstimmt (70a) wird diese als Antezedens gewählt. Zudem kann auch eine vorangehende NP als Antezedens ausgesucht werden, wenn der Prämodifikator der definiten NP mit dem Kopf der vorangehenden NP übereinstimmt (70b).

(70) a ***his art business*** *...* ***the art gallery***
b *a 15-acre plot and* ***main home*** *...* ***the home site***

[Vieira et al. 2000:560]

Bei der Auflösung von assoziativen Anaphern mit Hilfe von WordNet und Heuristiken haben [Poesio et al. 2002:1221] eine Ausbeute von 29% und eine Präzision von 38% verzeichnet. Es wurde aufgezeigt, dass assoziative Anaphern zwar mit Hilfe von WordNet angegangen werden können, dass sich der Erfolg solcher Ansätze allerdings in Grenzen hält, selbst wenn nur assoziative Anaphern in Meronymie- und Hyponymie-Beziehungen betrachtet werden.
WordNet birgt einige Nachteile, da viele Ausdrücke, Wortbedeutungen und lexikalische Beziehungen in WordNet nicht aufgeführt sind. So fehlt beispielsweise, wie [Poesio et al. 2002:1221] schreiben, der Eintrag für das Mineral „*crocidolite*“ und es ist zwar angegeben, dass *floor* Teil eines Raumes ist, aber nicht, dass ein Raum Teil eines Hauses ist. Durch den hierarchischen Aufbau von WordNet kann es vorkommen, dass die gewünschte Information nicht direkt ermittelt werden kann und auch wenn eine Relation zwischen einer definiten NP und einem möglichen Antezedenskandidaten in WordNet gefunden wird, muss es sich nicht um das korrekte Antezedens handeln. Ein weiteres Problem ist, dass Wörter mehrere Bedeutungen haben können und dies die Ursache von Fehlinterpretationen sein kann. Zudem ist der Aufbau von solchen Ressourcen teuer und zeitintensiv und es stellt sich die Frage, welche Information in solchen Quellen überhaupt enthalten sein muss. Sollte beispielsweise Alter mit Risikofaktoren in Zusammenhang gebracht werden?
Aufgrund der unbefriedigenden Ergebnisse mit WordNet wurden andere Ansätze zur Auflösung von assoziativen Anaphern getestet.

4.2.2.2.2 Auflösung von assoziativen Anaphern mit assoziativen Axiomen aus Korpora und WordNet

[Meyer et al. 2002:2] versuchen assoziative Beziehungen zwischen Nomen, wie *a bus* und *the driver* in Beispiel (71), aus einem Korpus zu extrahieren. Während diese Beziehungen in be-

stimmten Situationen nur implizit vorhanden sind (71a), können sie durchaus auch explizit sein (71b).

(71) a ***A bus*** *came around the corner.* ***The driver*** *had a mean look in her eyes.*
b *A bus nearly collided with a car.* ***The driver of the bus*** *had a mean look.*
[Meyer et al. 2002:1,2]

Das Ziel von [Meyer et al. 2002:3] ist es, aus einem Korpus assoziative Axiome auszufiltern und diese für die Auflösung von assoziativen Anaphern zu verwenden. Sie versuchen also explizit erwähnte Sachverhalte in dem Korpus zu erkennen. [Meyer et al. 2002:3] beschränken sich auf Genitivkonstruktionen und *of*-Konstruktionen, die oft auf assoziative Anaphern hindeuten. Aus dem Beispiel (72) kann das assoziative Axiom *have(stingray, head)* ermittelt werden.

(72) a ***The stingray's head*** *is not well defined, and there is no dorsal or caudal fin.*
b ***The head of the stingray*** *is not well defined, and there is no dorsal or caudal fin.*
[Meyer et al. 2002:3]

Mit den Konstruktionen „<NP of NP>“ und „<Genitive N>“ [Meyer et al. 2002:3] werden aber auch nicht-assoziative Konstruktionen gefunden, wie die Phrasen „*north of*“ und „*bird of prey*“. Es dürfen daher nur assoziative Fälle in Betracht gezogen werden.
[Meyer et al. 2002:3] benützen Hyperonymie-Beziehungen von WordNet, um Verallgemeinerungen aufzustellen. In WordNet kann beispielsweise die Information gefunden werden, dass „*heads*“ Körperteile und „*stingrays*“ Fische sind. Somit kann angenommen werden, dass Fische Körperteile haben und *the tuna* kann im Beispiel (73) als Antezedens von *the gills* gefunden werden.

(73) ***The tuna*** *has no respiratory mechanism to ensure the flow of water over* ***the gills****.*
[Meyer et al. 2002:3]

[Meyer et al. 2002:4] zeigen in Abbildung 26 auf, wie Verallgemeinerungen aufgestellt werden können. Das Hauptproblem dabei ist, dass Verallgemeinerungen keinesfalls immer zutreffen müssen. So kann aus dem Axiom in Abbildung 26[43] nicht geschlossen werden, dass Schlangen auch Beine haben.

[43] Die kursiven Einträge stellen das ursprüngliche Axiom dar; die kaptialisierten Einträge die Verallgemeinerungen.

```
have(okapi,leg)
have(okapi,LIMB)
have(GIRAFFE,leg)
have(GIRAFFE,LIMB)
…
have(LIVING THING, BODY PART)
```

Abb. 26 Axiome und Verallgemeinerungen von [Meyer et al. 2002:4]

[Meyer et al. 2002:5] haben getestet, wie das ursprüngliche Axiom und die aus den Verallgemeinerungen aufgestellten Axiome zum Finden des korrekten Antezedens einer assoziativen Anapher beitragen können. Das korrekte Antezedens soll aus einer Gruppe von möglichen Antezedenskandidaten herausgefiltert werden, indem die Axiome nacheinander auf die assoziative Anapher und die Kandidaten angewendet werden. In einem ersten Schritt wird nur das ursprüngliche Axiom angewendet, worauf hin die abgeleiteten Axiome ihrer Generalisationsstufe nach mit einbezogen werden. In jedem Schritt werden die Kandidaten ausgefiltert, für die kein Axiom zutrifft.

Je allgemeiner die Axiome sind, desto schlechter ist allerdings ihre Fähigkeit nicht zutreffende Kandidaten auszufiltern. Dafür haben zu spezifische Axiome die Tendenz korrekte Antezedenten auszuschliessen. Es bleibt also unklar, wie die Axiome sinnvoll als Filter für Antezedenskandidaten verwendet werden sollen und es stellt sich die Frage nach der Nützlichkeit dieses Ansatzes. Dennoch sind [Meyer et al. 2002:6] überzeugt, dass gekoppelt mit einem Gewichtungsansatz, die Axiome als zusätzlicher Faktor verwendet werden können, um den höchst wahrscheinlichen Kandidaten zu ermitteln.

4.2.2.2.3 Auflösung von assoziativen Anaphern mit fokusbasierten Techniken

[Poesio 2003:2] hat versucht, die Verarbeitung von assoziativen Anaphern mit fokusbasierten Techniken anzugehen. Sein Ansatz beruht auf der Centering-Theorie, die hier (basierend auf [Poesio 2003:4]) kurz erklärt wird. Gemäss der Centering-Theorie führt jede Äusserung neue Diskurseinheiten (Forward-Looking-Centers, CF) in den Diskurs ein. Der lokale Fokus wird aktualisiert, indem die neu eingeführten Einheiten eingeordnet werden. Es kann nach verschiedenen Kriterien geordnet werden. Wenn die neuen Einheiten nach ihrer grammatikalischen Funktion eingegliedert werden, werden beispielsweise Subjekte höher eingereiht als Objekte und Objekte höher als Adjunkte. Die Einordnung kann aber auch gemäss dem Informationsstatus der neuen Einheiten erfolgen; dem Leser bekannte Einheiten werden höher eingestuft als solche, die dem Leser noch nicht bekannt sind. Die sich in der Rangordnung zu-

oberst befindende Einheit wird als „Center“ bevorzugt (Preferred Center, CP). Das Backward-Looking Center (CB) entspricht dem Fokus. CB (U_i) der Äusserung U*i* ist das zuoberst eingereihte Element der CFs (U_{i-1}), das in U*i* realisiert wird.

Wenn nach dem Informationsstatus geordnet wird, haben 51.3% der assoziativen Anaphern, deren Antezedens in dem vorangehenden Satz zu liegen kommt, das CB(U_{-1}) als Bezugswort. Von allen assoziativen Anaphern, deren Antezedens nicht im selben Satz anzutreffen ist, sind es 33.6%. Werden wiederum auch intrasententielle assoziative Anaphern mitberücksichtigt, sind es knapp 29%. [Poesio 2003:6] hat ähnliche Ergebnisse erhalten, wenn nach grammatischen Funktionen geordnet wurde. Wird CP(U_{-1}) als Antezedens betrachtet, wurden leicht bessere Ergebnisse erzielt, aber nur wenn nach Informationsstatus geordnet wurde. Wie aus dieser Untersuchung klar wird, erhält man auch nicht viel versprechende Resultate, wenn man CBs oder CPs als Antezedens einer jeden assoziativen Anapher wählt. Wie [Poesio 2003:6] aber feststellen konnte, haben 89% der assoziativen Anaphern, die ihr Antezedens nicht im gleichen Satz haben, als Bezugswort eine Einheit, die vorher ein CB oder CP war. Dies sind immerhin ungefähr 60%, wenn alle assoziativen Anaphern mitberücksichtigt werden.

Es könnte sich also als vorteilhaft erweisen, das Antezedens einer assoziativen Anapher unter den vorherigen CPs und CBs zu suchen, vorausgesetzt man findet eine effektive Methode, um inkorrekte Antezedenskandidaten auszufiltern. [Poesio et al. 2004:3] haben dazu das World Wide Web (www), im Folgenden Web genannt, verwendet und mit dem Muster der Art „NBD of NPA“ versucht, die inkorrekten Antezedenskandidaten von assoziativen Anaphern auszufiltern. Dabei ist NBD das Hauptnomen der assoziativen Anapher und NPA das Hauptnomen des möglichen Antezedenskandidaten. 70% der assoziativen Anaphern konnten so richtig aufgelöst werden. Allerdings war das Testset sehr klein und bestand aus nur 20 assoziativen Anaphern.

4.2.2.2.4 Auflösung von assoziativen Anaphern mit Hilfe des Webs

[Bunescu 2003:2] versucht anhand von Abfragen im Web festzustellen, wie wahrscheinlich es ist, dass zwei Nomen in einer assoziativen Relation stehen. Anhand der Muster in Abbildung 27 retourniert eine Suchmaschine die mit der Suchanfrage übereinstimmenden Dokumente. Basierend auf der Anzahl der gefundenen Dokumente wird die Verbindung zwischen den zwei Nomen eingestuft.

$Q(n_t,n_a)$	=	„[n_t]. The [n_a][verb]“
$Q(n_t)$	=	„[n_t].“
$Q(n_a)$	=	„The [n_a][verb]“

Abb. 27 Suchmuster für assoziative Anaphern von [Bunescu 2003:3]

Während n_a für das Nomen der definiten NP steht, ist mit n_t das Antezedens der definiten NP gemeint. Das Muster $Q(n_t,n_a)$ in Abbildung 27 beschreibt die Situation, in der ein Satz mit n_t aufhört und der darauf folgende Satz mit einer definiten NP beginnt, die n_a enthält, und von einem der Verben *is/are*, *was/were*, *has/have*, *had*, *may*, *might*, *can*, *could*, *should*, *would* gefolgt wird. Eine Beispielsanfrage wäre „*soccer game. The players are*“.

Die Annahme ist, dass Identitätsanaphern und assoziative Anaphern und ihre Antezedenten in diesen Mustern ein hohes Ausmass an Assoziation aufweisen. Für die Auswahl dieses Musters spricht, dass viele Anapher-Antezedens Paare in einer solchen Konstruktion vorkommen, definite NPs in Identitätsanaphern und assoziativen Anaphern keine Modifikatoren, Relativsätze oder PPs benötigen, sowie dass die Verben sehr generell gewählt sind, damit jegliche Verbindung zwischen ihnen und den Nomen ausgeschlossen werden kann.

Zuerst wird die Anzahl der Dokumente ermittelt, die mit den Suchmustern in Abbildung 27 gefunden werden. Dann wird der Assoziationsgrad $I(n_t, n_a)$ berechnet, der die Abhängigkeit zwischen den Nomen n_t und n_a beschreibt. Die Berechnung ist in Abbildung 28 aufgezeigt.

N ist die Anzahl der von der Suchmaschine indexierten Dokumente. $D(query)$ ist die Anzahl der Webdokumente, die auf die Suchanfrage hin von der Suchmaschine (Altavista) ausgegeben werden.

$$I(n_t, n_a) = \log_2 \frac{P(Q(n_t, n_a))}{P(Q(n_t))\,P(Q(n_a))}$$

$$\approx \log_2 \frac{\frac{D(Q(n_t, n_a))}{N}}{\frac{D(Q(n_t))}{N} * \frac{D(Q(n_a))}{N}}$$

$$= \log_2 \frac{N*D(Q(n_t, n_a))}{D(Q(n_t)) * D(Q(n_a))}$$

Abb. 28 Berechnung des Assoziationswertes von [Bunescu 2003:3]

Es soll kurz verdeutlicht werden, was die Formel beinhaltet. Der Nenner des Bruchs ist die Wahrscheinlichkeit des gemeinsamen Auftretens der Einzelereignisse $Q(n_t)$ und $Q(n_a)$. Der Zähler ist die Wahrscheinlichkeit des Vorkommens des komplexen Ereignisses $Q(n_t, n_a)$, also

der Verknüpfung der Ereignisse $Q(n_t)$ und $Q(n_a)$. Wenn die Einzelereignisse unabhängig voneinander sind, tendieren der Nenner und der Zähler dazu, einen ähnlichen Wert zu haben und der Wert des Bruches liegt bei eins. Ist die Wahrscheinlichkeit des komplexen Ereignisses grösser als die Wahrscheinlichkeit des gemeinsamen Auftretens der Einzelereignisse, wird der Wert des Bruches grösser als eins. Da der Logarithmus von eins null ist, ergibt die Assoziation der Einzelereignisse Werte grösser null und die Disassoziation Werte kleiner null.

[Bunescu 2003:3,4] hat die Methode anhand von manuell ausgewählten definiten NPs getestet, für die ein Antezedens gefunden werden sollte. Dabei wurden nur assoziative definite NPs und definite NPs ohne NP-Antezedens evaluiert. Nicht-anaphorische definite NPs und Identitätsanaphern wurden nicht berücksichtigt. Die evaluierten definiten NPs beinhalteten zudem nur ein Nomen und hatten weder eine PP noch einen Relativsatz angehängt.

Da [Bunescu 2003:4] die 50 vorangehenden Nomen als mögliche Antezedenten für die Anapher ansieht, wurden 50 n_t: n_a Paare gebildet und ihr Assoziationswert berechnet. Als korrektes Antezedens wurde das Nomen n_t des n_t: n_a Paares angesehen, das den höchsten Assoziationswert erreichte. Ausserdem musste ein festgelegter minimaler Grenzwert erreicht werden, um von Assoziation sprechen zu können.

Der Ansatz von [Bunescu 2003] kann sich mit anderen Ansätzen, welche assoziative Anaphern behandeln, messen. Bei einer Ausbeute von 22.7% erreichte die Methode von [Bunescu 2003:5] eine Präzision von 53%, was heute auf dem Gebiet der automatischen Auflösung von assoziativen Anaphern akzeptabel ist.

[Bunescu 2003:5] hat die Fehler seines Ansatzes analysiert und festgestellt, dass nicht immer das Nomen des Paares mit dem höchsten Assoziationswert das korrekte Antezedens ist. Manchmal ist es das Nomen des Paares mit dem zweitbesten Assoziationswert. In diesen Fällen ist das korrekte Antezedens näher an der Anapher gelegen und somit muss die Distanz mitberücksichtigt werden. Ausserdem kann ein Nomen aufgrund der Anzahl der gefundenen Dokumente fälschlicherweise mit einem anderen assoziiert werden. Als Beispiel nennt [Bunescu 2003:5] die Assoziation von „*car:the butter*“, die anhand des Suchmusters gefunden wurde. Dies ist darauf zurückzuführen, dass mehrere Dokumente den gleichen Text enthalten und darum eine unsinnige Assoziation hoch gewertet wird. Eigentlich sollten Dokumente mit dem gleichen Text nur als ein Treffer gewertet werden.

Weiter kann die Assoziation zwischen Nomen komplexerer Art sein oder das korrekte Antezedens kann ausserhalb des Suchbereiches liegen, was zu fehlerhaften Assoziationen führen kann. Zudem können Änderungen des Diskursthemas Assoziationen zwischen Nomen ausser Kraft setzen, weshalb es von Vorteil wäre einen Algorithmus einzubauen, der Änderungen im

Diskursthema erkennt. Ein weiterer grosser Nachteil ist, dass der Ansatz von [Bunescu 2003:6] nur Nomen nimmt und keine NPs. Um auch NPs zu berücksichtigen muss ein Mechanismus zur Erkennung von Kollokationen eingebaut werden.

4.2.2.2.5 Auflösung von assoziativen Anaphern mit Ähnlichkeitskriterien

Der Ansatz von [Gasperin et al. 2004] zur Auflösung von assoziativen Anaphern soll hier kurz erläutert werden, obwohl er sich nicht auf das Englische bezieht. Statt WordNet verwenden [Gasperin et al. 2004:1] ein Ähnlichkeitsmass, um assoziative Anaphern im Portugiesischen und Französischen aufzulösen. Sie vergleichen dazu Nomen anhand ihrer Umgebung miteinander. Die Nomen sind sich umso ähnlicher, je mehr syntaktischen Kontext sie gemeinsam haben. Als Beispiel eines solchen syntaktischen Kontexts nennen [Gasperin et al. 2004:1] die Beziehung zwischen dem Subjekt und dem Verb. So haben zwei Nomen, die beide Subjekt eines bestimmten Verbs sein können, einen gemeinsamen Kontext. Die generierten Ähnlichkeitslisten enthalten Nomen, welche in verschiedenen lexikalischen Beziehungen, wie zum Beispiel Synonymie, Hyponymie und Meronymie zu dem aktuellen Nomen stehen.

Wie [Gasperin et al. 2004:2] schreiben, sind sich beispielsweise die Wörter „*accusation*“ und „*denunciation*“ sowie die Wörter „*boss*“ und „*employers*“ ähnlich. Es werden aber auch Wörter als ähnlich befunden, die einfach dem gleichen semantischen Bereich oder Typ angehören.

Der Vorteil dieses Ansatzes gegenüber WordNet ist, dass nur semantische Beziehungen in Betracht gezogen werden und Ambiguitäten nicht berücksichtigt werden müssen[44]. Es können Verbindungen zwischen Wörtern ermittelt werden, die aus einem Wörterbuch nicht ersichtlich sind. [Gasperin et al. 2004:3] nennt hier die Ähnlichkeit des Nomens „*micro*“ wie in „*micro computer*“ zu „*386 DX, 40MHz, 8 Mbytes de RAM*“.

Wie wird nun anhand solcher Ähnlichkeitslisten das Antezedens ermittelt? Das Antezedens der NP-Anapher wird innerhalb der Liste der ähnlichen Nomen vermutet. Oder aber es wird in der Ähnlichkeitsliste jedes Antezedenskandidaten nach der Anapher gesucht. Ausserdem kann auch geprüft werden, ob die Hauptnomen der Anapher und des Antezedens in der gleichen Ähnlichkeitsliste vorkommen oder ob es ein drittes Wort gibt, dass der Anapher sowie dem Antezedens ähnlich ist. Die Resultate dieses Ansatzes lassen sich allerdings nur schwer

[44] Dem System muss nicht bekannt sein, dass das Wort „Bank“ mehrere Bedeutungen hat (Finanzinstitut, Sitzbank). Wenn beispielsweise die Wörter „Bank“ und „Institution“ in einem Text vorkommen, muss das System nur eine Verbindung zwischen diesen Wörtern herstellen können; die andere Bedeutung von Bank (Sitzbank) kann ausser Acht gelassen werden.

mit den anderen ähnlichen Ansätzen vergleichen, da sie auf das Portugiesische und Französische ausgerichtet sind. Aus der Evaluation des Ansatzes geht aber hervor, dass nicht schlechtere Resultate als mit WordNet erzielt werden.

4.2.2.2.6 Auflösung von assoziativen Anaphern mit vektorbasierten Methoden

[Poesio et al. 2002:1222] haben vektorbasierte Methoden zur Auflösung von assoziativen Anaphern getestet. Diesen Methoden liegt die Annahme zugrunde, dass eine Bedeutung eines Wortes als Vektor mit verschiedenen Merkmalen dargestellt werden kann. Solche Merkmale können beispielsweise andere Wörter sein, die in der Umgebung der assoziativen Anapher auftreten. Gemäss [Poesio et al. 2002:1223] treten assoziative Anaphern in Meronymie-Beziehungen häufig in syntaktischen Konstruktionen wie „*the* NP *of* NP“, „NP *of* NP“, „NP*'s* NP“ auf. Anhand von Testkorpora konstruiert der Algorithmus für jedes Hauptnomen *N* einer assoziativen Anapher einen Vektor, der die Häufigkeit der Vorkommen von anderen Wörtern zusammen mit *N* in einer der oben genannten Konstruktionen erfasst.
Das wahrscheinlichste Antezedens aus einer Auswahl von möglichen Bezugswörtern innerhalb der fünf vorangehenden Sätze wird mit der Formel in Abbildung 29 berechnet, die der Formel in Abbildung 28 entspricht.

$$I(x;y) = \log \frac{P(x,y)}{P(x)\,P(y)}$$

Abb. 29 Berechnung des Bezugswortes in Meronymie-Beziehungen von [Poesio et al. 2002:1223]

Mit dieser Methode konnten [Poesio et al. 2002:1223] die Auflösung von assoziativen Anaphern gegenüber dem WordNet-Ansatz deutlich verbessern. Für die Auflösung von assoziativen Anaphern in einer Meronymie-Beziehung verzeichneten sie eine Ausbeute von 66.7% und eine Präzision von 72.7%.

4.2.2.2.7 Auflösung von assoziativen Anaphern mit Techniken des ML

[Poesio et al. 2004:1] verwenden Techniken des MLs, um assoziativen Anaphern in Meronymie-Beziehungen, also Teil-Ganzes-Beziehungen, aufzulösen. Dabei spielen der lokale Fokus und die lexikalische Distanz eine bedeutende Rolle. Wie [Poesio 2003:5] ermittelt haben, be-

vorzugen assoziative Anaphern Einheiten in demselben und dem vorangehenden Satz als Antezedens, wie Abbildung 30 zeigt.

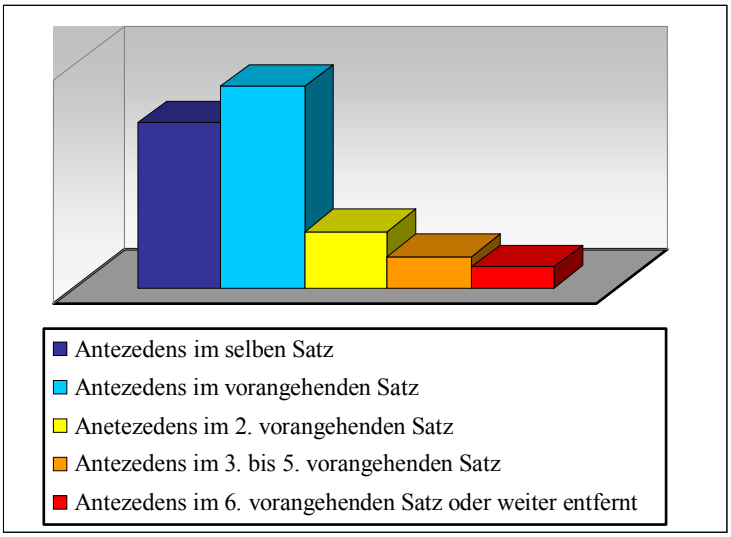

Abb. 30 Distanz zwischen assoziativen Anaphern und ihrem Antezedens, basierend auf [Poesio 2003:5]

Die lexikalische Distanz wird von [Poesio et al. 2004:4] einerseits mit der Suchmaschine Google aus dem Web und andererseits aus WordNet berechnet. Abbildung 31 zeigt, wie die lexikalische Distanz aus Google ermittelt wird.

$$\text{Google distance} = \begin{cases} 1 & \text{if } NHits = 0 \\ \dfrac{1}{NHits} & \text{otherwise} \end{cases}$$

Abb. 31 Google-Distanz von [Poesio et al. 2004:4]

Es wird mit dem Suchmuster „*the* NBD *of the* NPA“[45] gesucht, wobei NBD das Hauptnomen der assoziativen Anapher ist und NPA das Hauptnomen des Antezedenskandidaten. *NHits* ist die Anzahl der Treffer, die Google mit der Suchanfrage findet.

Die lexikalische Distanz aus WordNet wird ermittelt, indem geprüft wird, ob eine Hyperonymie-Beziehung zwischen der assoziativen Anapher und dem Antezedens vorherrscht. Die Distanz(*s,s'*) ist die Anzahl von Hyperonymie-Verbindungen zwischen den Konzepten *s* und *s'*. Zuerst werden von WordNet alle Bedeutungen von NBD und NPA und die dazugehörenden Hyperonymie-Bäume herausgesucht. Für jedes Bedeutungspaar s_{NBDi} und s_{NPAj} wird das

[45] Wichtig ist hier, dass die Anfrage in Anführungszeichen steht, d.h. NBD und NBA nicht durch bool-sche Operatoren verknüpft sind.

spezifischste Konzept s_{comij}, das ein Hyperonym beider Bedeutungen ist, herausgesucht. Die kürzeste WordNet-Distanz (*ShtstWNDist*) zwischen NBD und NPA wird als kürzeste Distanz zwischen den Bedeutungen von NBD und NPA betrachtet, wie Abbildung 32 zeigt.

$$ShtstWNDist(\text{NBD},\text{NPA}) = \min_{i,j}(\text{distance}(s_{NBDi}, s_{commij}) + \text{distance}(s_{comij}, s_{NPAj}))$$

Abb. 32 Kürzeste WordNet-Distanz von [Poesio et al. 2004:5]

Eine normalisierte WordNet-Distanz zwischen 0 und 1 wird erreicht, wenn die kürzeste WordNet-Distanz mit einem Faktor *MaxWNDist* geteilt wird, der bei [Poesio et al. 2004:5] den Wert 30 hat. Die WordNet-Distanz ist 1, falls keine Verbindung zwischen den Konzepten gefunden werden kann.

$$\text{WN distance} = \begin{cases} 1 & \text{if no path} \\ \dfrac{ShtstWNDist}{MaxWNDist} & \text{otherwise} \end{cases}$$

Abb. 33 WordNet-Distanz von [Poesio et al. 2004:5]

Neben der lexikalischen Distanz werden auch andere Merkmale verwendet. Die Äusserungsdistanz gibt die Entfernung zwischen der Äusserung an, welche die Anapher enthält und derjenigen Äusserung, die den Antezedenskandidaten beinhaltet. Die Äusserung wird mit dem Satz gleichgesetzt. 81% der von [Poesio et al. 2004:5] untersuchten assoziativen Anaphern in Teil-Ganzes-Beziehungen hatten ihr Antezedens in den vorangehenden zwei Sätzen. Kein Antezedens war jedoch im selben Satz wie die Anapher gelegen.

Auch die ersterwähnte Einheit in einem Satz wird als Merkmal betrachtet, da assoziative Anaphern ersterwähnte Einheiten als Antezedens zu bevorzugen scheinen. Weit seltener ist die letzterwähnte Einheit im vorangehenden Satz das korrekte Bezugswort, was darauf hindeutet, dass zuerst erwähnte Einheiten über den zur assoziativen Anapher nächstgelegenen als Antezedens bevorzugt werden. Die ersterwähnte Einheit wurde im lokalen sowie im globalen Kontext ermittelt. Lokal wurden die fünf vorangehenden Sätze betrachtet, global wurde der gesamte Kontext mitberücksichtigt. 53.5% der Antezedenten von den Anaphern in Teil-Ganzes-Beziehungen waren als ersterwähnte Einheit im lokalen und 56.9% im globalen Kontext realisiert.

Das Testset bestand aus positiven Beispielen, die eine korrekte Verbindung zwischen einer Anapher und ihrem Antezedens signalisierten und negativen Beispielen, die inkorrekte Ver-

bindungen beinhalteten. Es wurden auch indefinite NP-Anaphern miteinbezogen. Die durchgeführten Evaluationsexperimente sind in [Poesio et al. 2004:6] nachzulesen. Kurz zusammengefasst verdeutlichen die Experimente, dass mit der Kombination von verschiedenen Merkmalen bessere Ergebnisse erzielt werden. Mit der WordNet-Distanz und der Google-Distanz erhält man ähnliche Resultate. Die besten Ergebnisse werden erreicht, wenn sowohl die Google-Distanz als auch die WordNet-Distanz verwendet werden.

Es wurden einige Ansätze zur Auflösung von assoziativen NP-Anaphern betrachtet. Ersichtlich ist, dass die automatische Auflösung von assoziativen Anaphern erst in den Kinderschuhen steckt und die Ansätze noch nicht ausgereift sind. Zuversichtlich stimmt allerdings, dass sich immer mehr Forscher mit anaphorischen Beziehungen komplexer Art auseinandersetzen.

4.2.2.3 Automatische Auflösung von *Other*-Anaphern

Other-Anaphern sind definite NPs, die mit *other* oder *another* beginnen. *Other*-Anaphern stehen in einer ergänzenden Relation zu ihrem Antezedens. Im Beispiel (74) verweist die NP *other risk factors for Mr. Cray's company* auf eine Gruppe von Risikofaktoren, in der allerdings der Risikofaktor Alter nicht enthalten ist. Anzumerken ist, dass die Distanz zwischen einer *Other*-Anapher und ihrem Antezedens beträchtlich sein kann.

(74) *You either believe Seymour can do it again or you don't. Beside* ***the designer's age, other risk factors*** *for Mr. Cray's company include the Cray-3's chip technology.*

[Markert et al. 2003:1]

Für die Auflösung von *Other*-Anaphern ist lexikalisches Wissen und Weltwissen unentbehrlich. [Markert et al. 2003:2] stellen ein System zur Auflösung von *Other*-Anaphern vor, welches das Web als Ressource verwendet. Wie sie schreiben, muss beispielsweise Alter nicht als Risikofaktor in einer Ontologie verankert werden. Vielmehr muss ein System entscheiden können, welche der NPs (*Seymour*, *designer*, *designer's age*) im Beispiel (74) einen Risikofaktor darstellen kann.

[Markert et al. 2003:2] verwenden Suchmuster und Web-Frequenzen, um das korrekte Antezedens aus möglichen Kandidaten auszuwählen. Zwischen einer *Other*-Anapher und ihrem Antezedens herrscht oft eine Hyponymie- oder Ähnlichkeitsbeziehung vor. Sie wählen daher Suchmuster aus, die solche lexikalischen Beziehungen ausdrücken. Die Suchmuster sind in der Abbildung 34 aufgeführt.

Antezedens	**Muster**	**Instantiierung**	
Nomen	(N_1\{sg\}OR N_1\{pl\}) and other N_2 \{pl\}	(1)	(age OR ages) and other factors
Namen	(N_1\{sg\}OR N_1\{pl\}) and other N_2\{pl\}) N_1 and other N_2	(2) (3) (4)	(person OR persons) and other shareholders (shareholder OR shareholders) and other persons Mr. Pickens and other shareholders

Abb. 34 Suchmuster für *Other*-Anaphern im Web-Ansatz von [Markert et al. 2003:5]

Namen müssen in den Suchmustern durch allgemeine Kategorien wie Personen, Organisationen, Orte und Flüsse ersetzt werden, um ein vernünftiges Suchresultat zu erhalten. Auf die Klassifizierung von Namen durch bestimmte Namenerkennungs-Systeme soll hier aber nicht näher eingegangen werden.

Anhand des Suchmusters werden die Häufigkeiten ermittelt, in denen die *Other*-Anapher mit ihren möglichen Antezedenten in Verbindung steht. Für Beispiel (74) bedeutet dies, dass nach „*Seymour and other risk factors*“, nach „*designer and other risk factors*“ und nach „*age and other risk factors*“ [Markert et al. 2003:2] gesucht wird. Als Antezedens wird die mit der *Other*-Anapher am häufigsten auftretende NP gewählt. Während für die beiden ersten Instantiierungen kein Eintrag mit der Suchmaschine Google gefunden wird, ergibt die Suche für *age and other risk factors* laut [Markert et al. 2003:3] über 400 Suchergebnisse.

Für Nomen wird die Häufigkeit $freq((1))$ und für Namen die Häufigkeiten $freq((2))$, $freq((3))$ und $freq((4))$ berechnet (siehe Abbildung 34 und 35). Es wird nun das Maximum M_{ant} über diesen Frequenzen ermittelt. Für gewöhnliche Nomen entspricht das Maximum $freq((1))$. Die Instantiierung, die dann M_{ant} erreicht, wird $Imax_{ant}$ genannt.

$$Pr(Imax_{ant}) = \frac{M_{ant}}{number\ of\ Google\ pages}$$

$$Pr(X_{ant}) = \frac{freq(X_{ant})}{number\ of\ Google\ pages}$$

$$Pr(Y_{ant}) = \frac{freq(Y_{ant})}{number\ of\ Google\ pages}$$

$$MI_{ant} = \log \frac{Pr(Imax_{ant})}{Pr(X_{ant})\ Pr(Y_{ant})}$$

Abb. 35 Häufigkeitsberechnung der Vorkommen von Anapher-Antezedens-Paaren in Google von [Markert et al. 2003:5]

Es gibt zwei Bewertungsmethoden. In der ersten Methode wählen [Markert et al. 2003:4] den Kandidaten als korrektes Antezedens aus, der das höchste Maximum M_{ant} erreicht. In der zweiten Methode werden die individuellen Häufigkeiten von den Antezedenskandidaten und

der Anapher berücksichtigt. Der erste Teil von $Imax_{ant}$, der beispielsweise „*age* OR *ages*“ oder „*shareholder* OR *shareholders*“ wird von [Markert et al. 2003:4] X_{ant} genannt, der zweite Teil („*factors*“ oder „*persons*“) ist Y_{ant}. Es wird die Wahrscheinlichkeit *Pr* von $Imax_{ant}$, X_{ant}, Y_{ant} berechnet, wobei Google verwendet wird, um die Häufigkeiten $freq(X_{ant})$ und $freq(Y_{ant})$ zu bestimmen. Der Kandidat mit der höchsten MI_{ant} wird als Antezedens ausgewählt. Wenn zwei Kandidaten den gleichen Wert erhalten, wird derjenige Kandidat als Antezedens gewählt, welcher der Anapher näher ist.

Wie aus der Abbildung 36 ersichtlich ist, lassen sich die Resultate mit Methoden vergleichen, die WordNet verwenden. Ein solches System ist LEX von [Modjeska 2002, 2003]. *Other*-Anaphern, die das gleiche Hauptnomen aufweisen wie ihr Antezedens (*90 days ... another 90 days* [Modjeska 2003:64]), werden ermittelt, indem das Hauptnomen der Anapher mit den Hauptnomen der übrigen NPs in einem bestimmten Suchbereich verglichen wird. Als Antezedens wird die NP gewählt, die das gleiche Hauptnomen wie die Anapher aufweist. *Other*-Anaphern, die in einer Hypernymie-Beziehung zu ihrem Antezedens stehen (*cuckoo's habit… other birds* [Modjeska 2002:130]), werden anhand von Hypernymie-Bäumen aus WordNet gefunden. Dabei wird der Baum des Antezedens mit den Hypernymie-Bäumen der Antzedenskandidaten verglichen. Als Antezedens wird die NP ausgegeben, die ein direktes oder indirektes Hyperonym der Anapher ist.

***Other*-Anaphern**	**Methode 1**	**Methode 2**	**LEX**
korrekt aufgelöst	57	63	61
falsch aufgelöst	63	57	59

Abb. 36 Auflistung von korrekt und falsch aufgelösten *Other*-Anaphern im webbasierten Ansatz von Markert 2003] und dem WordNet-Ansatz LEX, basierend auf [Markert et al. 2003:5]

Die Fehler sind der Nichterkennung oder Falschklassifikation von Namen, vagen Anaphern wie „*issue*“ oder „*problem*“ [Markert et al. 2003:5] sowie Pronomen und geteilten Antezedenten, mit denen das System nicht umgehen kann, zuzuschreiben.

[Markert et al. 2003: 6] haben ihre Methode zur Auflösung von *Other*-Anaphern auch verwendet, um assoziative Anaphern in Meronymie-Beziehungen (75) anzugehen.

(75) ***The apartment*** *she shares with a 12-year-old daughter and her sister was rattled, books and crystal hit* ***the floor****…*

[Markert et al. 2003:1]

[Markert et al. 2003:6] haben die Suchmuster der Abbildung 37 verwendet und wiederum akzeptable Ergebnisse erhalten, die sich mit WordNet-Ansätzen vergleichen lassen. Allerdings war das Testset mit nur einem Dutzend assoziativen Anaphern sehr klein und enthielt keine komplexeren Fälle, weshalb die Daten nicht repräsentativ sind.

Muster	**Instantiierung**
(N_1 {sg}OR N_1 {pl}) of (a OR an OR the OR each OR every OR any)* N_2 {sg}	(floor OR floors) of (a OR an OR the OR each OR every OR any)*apartment
(N_1 {pl}of (the OR all)* N_2 {pl})	floors of (the OR all)*apartments

Abb. 37 Suchmuster für assoziative Anaphern in Meronymie-Beziehungen von [Markert et al. 2003:6]

Neben *Other*-Anaphern gibt es weitere Anaphern, die mit komparativen Modifikatoren (z.Bsp. *such*, *similar*, *same*) beginnen und auch bei der automatischen Behandlung von Anaphern mitberücksichtigt werden müssten. Ansätze zur automatischen Auflösung von solchen Anaphern wurden bis jetzt aber noch nicht entwickelt.

4.2.2.4 Ein Algorithmus zur Erkennung von anaphorischen NPs

Wie bereits erläutert wurde, müssen auch anaphorische NPs in einem Text erkannt werden. [Cardie et al. 2002:1] haben eine Methode zur Bestimmung von anaphorischen NPs entwickelt, die als Filter in ein AR-System bzw. Koreferenzauflösungssystem eingebaut werden kann. Der Ansatz hat den Vorteil, dass keine handgeschriebenen Heuristiken nötig sind, dass er genreunabhängig ist und dass alle Typen von NPs, also auch indefinite NPs, behandelt werden können.

Die Komponente zur Bestimmung von anaphorischen NPs kann mit Methoden des MLs trainiert werden. [Cardie et al. 2002:2] verwenden dazu Daten aus koreferenzannotierten Korpora. Jede Einheit setzt sich aus einer einzelnen NP zusammen und beinhaltet 37 Merkmale, die lexikalischer, grammatikalischer, semantischer und positionaler Art sind. Anhand dieser Merkmale kann eine NP als anaphorisch beziehungsweise nicht-anaphorisch klassifiziert werden. Ein Merkmal prüft beispielsweise, ob eine NP *NPi* existiert, die einer *NPj* vorangeht und *NPi* und ob *NPj* denselben Kopf haben. Ist dies der Fall, wird der NP *NPi* ein bestimmter Wert zugeordnet. Diese Merkmalswerte werden automatisch berechnet. Die Merkmale sollen hier nicht näher erläutert werden, eine detaillierte Auflistung ist in [Cardie et al. 2002:4] zu finden.

Falschklassifikationen können zu Einbussen im eigentlichen System führen. Als Abhilfe kann der Algorithmus zur Bestimmung von anaphorischen und nicht-anaphorischen NPs übergangen werden, wenn sich zwei NPs aus den gleichen Wortketten zusammensetzen oder es sich um Synonyme handelt.

Nachdem nun einige Aspekte der automatischen Behandlung von vollständigen NPs behandelt worden sind, sollen nun zwei konkrete Systeme vorgestellt werden.

4.2.3 AR-Systeme zur Behandlung von vollständigen NPs

4.2.3.1 Der Ansatz von [Vieira et al. 2000]

[Vieira et al. 2000] haben einen Ansatz zur Behandlung von anaphorischen und nicht-anaphorischen vollständigen NPs entwickelt. Ihr System versucht zwischen nicht-anaphorischen definiten NPs, Identitäts-Anaphern und assoziativen Anaphern zu unterscheiden. Abbildung 38 zeigt eine Darstellung dieses Systems.

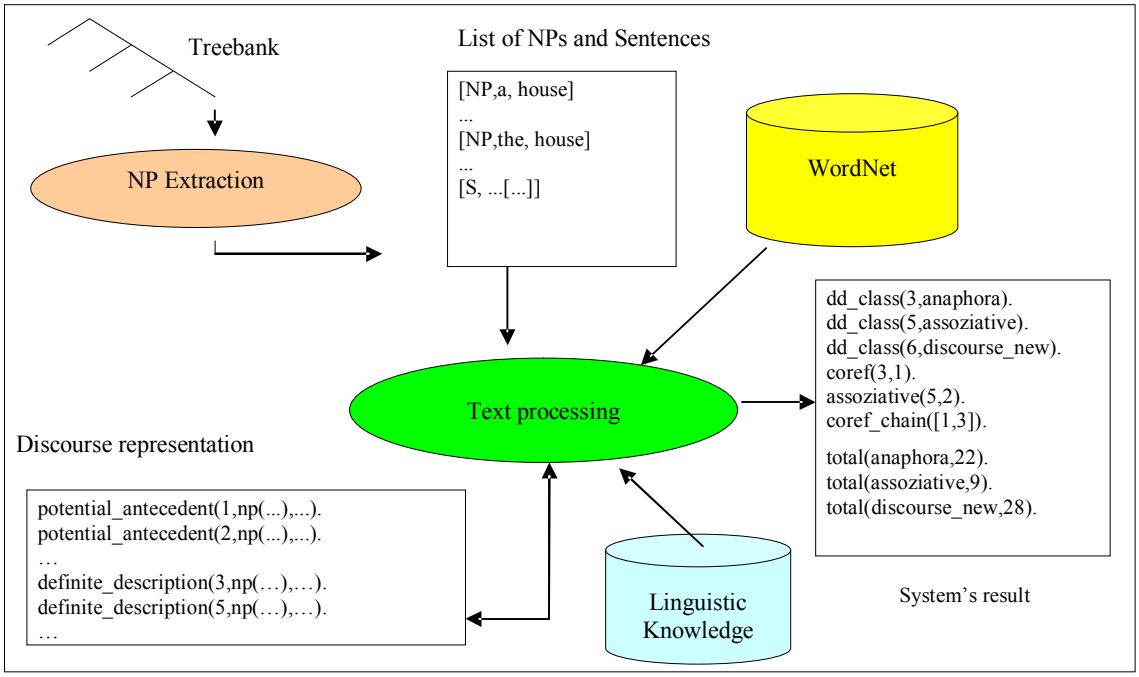

Abb. 38 Darstellung des Systems von [Vieira et al. 2000:545] zur Behandlung von definiten NPs

Die Sätze aus der Treebank[46] werden als Listen dargestellt und die NPs extrahiert. Das System klassifiziert jede definite NP und findet ein Antezedens für anaphorische NP-Anaphern.

[46] Eine Treebank ist ein Korpus mit bereits geparsten Sätzen

Die Klassifikation der definiten NPs basiert auf Heuristiken, die bereits in 4.2.1 und 4.2.2 beschrieben worden sind. Die Ablaufstrategie ist in Abbildung 39 aufgezeigt.

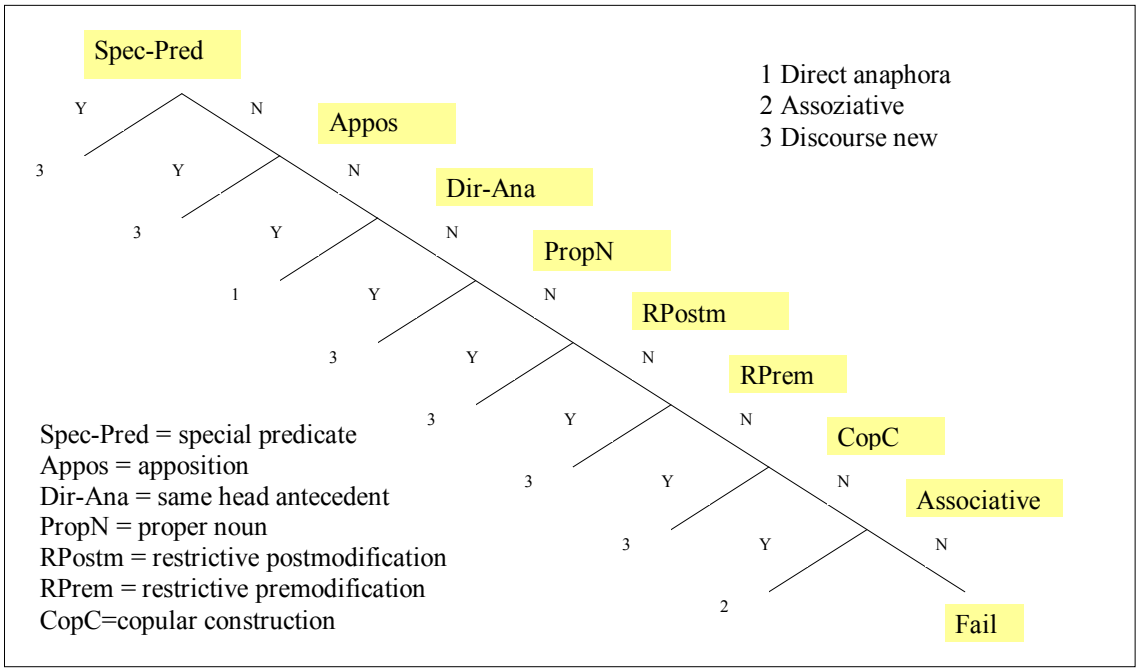

Abb. 39 Entscheidungsbaum zur Klassifikation von definiten NPs von [Vieira et al. 2000:577]

Zuerst wird versucht die definite NP anhand von Heuristiken als nicht-anaphorisch zu klassifizieren. Wenn dies nicht gelingt, wird nach einem Antezedens mit demselben Kopf gesucht. Wenn kein solches gefunden werden kann, wird geprüft, ob die definite NP diskursneu ist. Erst danach wird untersucht, ob die definite NP als assoziative NP interpretiert werden kann.
Die Heuristiken zur Erkennung von assoziativen Anaphern werden erst angewendet, wenn alle übrigen Heuristiken versagen. Dies liegt einerseits daran, dass die Performanz dieser Heuristiken eher schlecht ist, und die Berechnung von assoziativen NPs zeitintensiv ist. Wenn keine Heuristik zutrifft, wird der definiten NP keine Kategorie zugeordnet.
[Vieira et al. 2000:581] haben zwei Evaluationen des Systems durchgeführt. Im Gegensatz zu der Version 1, wird in der Version 2 auch versucht, assoziative NP-Anaphern aufzulösen. Die Resultate sind in Abbildung 40 aufgezeigt.

Version	***Recall***	***Precision***	***F-measure***
Version 1	53%	76%	62%
Version 2	57%	70%	62%

Abb. 40 Zwei Evaluationsversionen des AR-Systems von [Vieira et al. 2000:581]

Fehleranalysen des Systems haben ergeben, dass die Modifikation nicht ganz unproblematisch ist. Obwohl unterschiedliche Prämodifikatoren üblicherweise nicht auf Koreferenz schliessen lassen, gibt es Ausnahmen (76a). Ausserdem kann der Algorithmus auch definite NPs als anaphorisch klassifizieren, obwohl zwischen ihnen keine solche Relation vorherrscht, wie in (76b). Ausnahmen gibt es auch mit definiten NPs, die von Eigennamen modifiziert werden und normalerweise nicht-anaphorisch sind (76c). Auch kann das System nicht-anaphorische definite NPs wie *the world, the nation* etc. [Vieira et al. 2000:577] nicht als diskursneu erkennen.

(76) a *the pixie-like clarinettist… **the soft-spoken clarinettist***
b *the population… the voting population*
c *Others grab books, records, photo albums, sofas and chairs, working frantically in the fear that an aftershock will jolt **the house** again. …*
*As Ms. Johnson stands outside **the Hammock** house after winding up her chores there, the house begins to creak and sway.*
[Vieira et al. 2000:566,549]

Fehler sind auch auf die Rechtschreibung und den beschränkten Suchbereich zurückzuführen. Zudem sollten neben indefiniten, definiten und possessiven NPs auch andere Einheiten als Antezedenten zugänglich sein. Wenn das System kein Antezedens für eine definite NP findet, klassifiziert es diese als nicht-anaphorisch, was nicht immer richtig sein muss. In einigen Fällen klassifiziert das System eine definite NP korrekterweise als anaphorisch, ordnet ihr aber ein falsches Antezedens zu. Das System kann ausserdem nicht mit Numerusinkongruenz zwischen Anapher und Antezedens umgehen. Weiter hat der Algorithmus Mühe mit zusammengesetzten NPs, die mehr als zwei Nomen enthalten, da diese die gleiche Struktur wie Appositionen haben. Zahlreiche Fehlinterpretationen sind auch auf den Gebrauch von WordNet zurückzuführen.

4.2.3.2 Der Ansatz von [Postolache 2004]

Die von [Postolache 2004:8-11] aufgeführten Schwierigkeiten in Kapitel 4.2.2 werden von [Postolache 2004] sowie von [Cristea et al. 2005] mit einem robusten AR-System namens RARE (Robust Anaphora Resolution Engine) angegangen, weshalb dieses System nun beschrieben werden soll. Die Beschreibung basiert auf [Postolache 2004:13-37] und [Cristea et al. 2005:18-28]. Das System berücksichtigt sowohl linguistische als auch semantische Einheiten, die im Prozess der AR von Bedeutung sind. Dabei werden linguistischen Einheiten wie Pronomen, NPs und Zahlen als referentielle Ausdrücke (REs) in einem Text-Layer

dargestellt. Die semantischen Einheiten[47] werden als Diskurseinheiten (DEs) in einem semantischen Layer aufgeführt. Die DEs beinhalten bestimmte Merkmale der REs, auf die unten näher eingegangen wird. Im Projektions-Layer werden Informationen der REs, d.h. bestimmte Merkmale der REs, in so genannte Projektionsstrukturen (PSs) gefüllt. Der Auflösungsprozess beruht auf der Übereinstimmung der PSs und der DEs.

Wenn unterschiedliche REs auf ein und dieselbe DE verweisen, handelt es sich um koreferente Ausdrücke. Eine solche Beziehung ist in Abbildung 41 dargestellt. Abbildung 42 zeigt eine assoziative Relation im RARE-System. Beide Beziehungen sind anaphorisch.

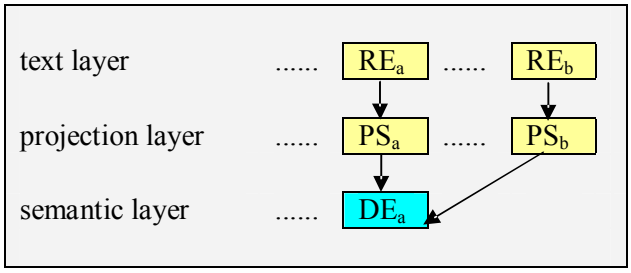

Abb. 41 Darstellung einer Koreferenz-Relation im RARE-System, basierend auf [Postolache 2004:14]

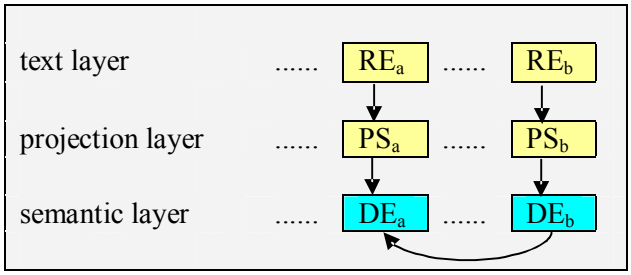

Abb. 42 Darstellung einer assoziativen Relation im RARE-System, basierend auf [Cristea 2005:20]

Wenn eine RE anhand der PS einer DE zugeordnet werden kann, gibt es eine direkte bidirektionale Verbindung von der RE zur DE und die PS kann gestrichen werden. Wenn die RE über die PS keiner bereits bestehenden DE zugeordnet werden kann, wird eine neue DE geschaffen, indem die Merkmale der PS auf die DE kopiert werden. Dabei gibt es wiederum eine direkte Verbindung zwischen der RE und der DE. Siehe dazu Abbildung 43.

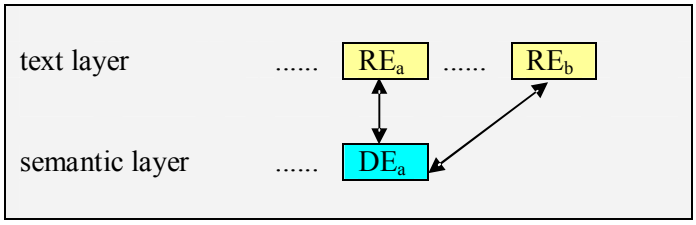

Abb. 43 Darstellung einer anaphorischen Relation im RARE-System, basierend auf [Postolache 2004:14]

Im RARE-System können unterschiedliche Modelle zur AR untergebracht sein. Ein solches Modell beinhaltet verschiedene Komponenten: Einerseits beinhaltet das System Attribute der

[47] Mit semantischen Einheiten ist die Repräsentation der Einheiten gemeint, von denen ein Diskurs handelt. Im Beispiel „*John bought an apple. He ate it.*“ gibt es zwei semantische Einheiten (John, Apfel). Die REs dieses Satzes sind *John*, *an apple*, *he* und *it*. Jede Erwähnung einer semantischen Einheit wird also als RE dargestellt.

PSs und DEs, d.h. Merkmale morphologischer, syntaktischer, lexikalisch-semantischer und positionaler Art. Andererseits sind auch Wissensquellen, Matching-Regeln und Heuristiken enthalten. Mit den Heuristiken kann bestimmt werden, ob eine PS eine neue DE hervorruft oder auf welche bestehende DE sie sich bezieht. Weiter gibt es auch Heuristiken, die den Verarbeitungsprozess konfigurieren und beispielsweise die Verarbeitungsreihenfolge der DEs und Suchbereichkriterien festlegen.

Mögliche REs sind im System von [Postolache 2004:23] NPs, Pronomen und Zahlen. Zu den NPs gehören definite, indefinite und unbestimmte NPs sowie Eigennamen, Datumsangaben, Währungsausdrücke und Prozentangaben. Unter Pronomen sind Personal-, Possessiv-, Reflexiv-, Demonstrativ- und Relativpronomen[48] zusammengefasst. Zahlen werden als REs markiert, wenn sie sich auf Einheiten beziehen („*four of them*“ [Postolache 2004:24]), nicht aber wenn sie die Funktion eines Adjektivs oder Adverbs haben („*three books*“).

Als Attribute hat [Postolache 2004:24] Lemma (lemma[49]), Numerus (number), Wortart (pos), syntaktische Funktion (role[50]), Name (femaleName, maleName, familyName[51]), Apposition (apposition[52]), Synonym (synonyms[53]) und Hypernym (hypernyms[54]) gewählt. Neben diesen Attributen wird auch ein Merkmal HeSheItThey aufgeführt, welches die Wahrscheinlichkeit beschreibt, mit der auf eine NP durch die Pronomen *he*, *she*, *it* oder *they* verwiesen wird. Das Attribut includes[55] beschreibt REs, die in andere REs eingebettet sind. Das Attribut indefinite[56] bestimmt, ob es sich um indefinite NPs handelt. Weiter gibt es ein Attribut wh, dessen Wert „yes“ ist, wenn es sich bei den REs um „*which*, *who*, *whose*, *whom* und *that* handelt.

Für jedes Attribut wurden Wissensquellen mit den entsprechenden Werten definiert. Solche Wissensquellen sind beispielsweise POS-Tagger, FDG-Parser, Namenslisten, und WordNet. Als weitere Wissensquelle hat [Postolache 2004:25] einen Algorithmus zur Erkennung von Appositionen eingebaut. Die automatische Erkennung von Appositionen ist eine schwierige Aufgabe, da die Zeichensetzung kein verlässlicher Hinweis ist. Das System von [Postolache 2004:25] verwendet deshalb lexikalische, syntaktische und semantische Einschränkungen

[48] Relativpronomen, die in Fragen vorkommen werden nicht als REs angesehen (z.Bsp. *Who is your boss?*)
[49] Lemma des Kopfes der RE
[50] Syntaktische Funktion des Kopfes der RE
[51] femaleName hat den Wert „yes“, wenn es sich beim Lemma um einen weiblichen Namen handelt. (Entsprechend: maleName bei männlichen Namen und familyName bei Familiennamen)
[52] Das Attribut beinhaltet die Identifikationsnummer (ID) der RE, die in einer appositiven Verbindung zur aktuellen RE steht
[53] Das Attribut beinhaltet die Liste mit den WordNet-Synonymen des Lemmas der RE
[54] Das Attribut beinhaltet die Liste mit den WordNet-Hypernymen
[55] Das Attribut beinhaltet die Identifikationsnummern (IDs) von REs, die in der aktuellen RE eingebettet sind. Das Attribut kann daher auch leer sein.
[56] Das Attribut hat den Wert „yes“, wenn die RE eine indefinite NP ist und „no“, wenn sie eine definite NP ist

beim Finden von Appositionen. Ein Beispiel einer Einschränkung ist die Numeruskongruenz, die eingehalten werden muss und somit Beispiele wie (77) nicht als Apposition ansieht.

(77) *He didn't trust her smile, her eyes.*

[Postolache 2004:25]

Eine spezielle Wissensquelle ist diejenige, der die Werte für das HeSheItThey Attribut entnommen werden können. [Postolache 2004:26] verwenden Wahrscheinlichkeitswerte für die Zuordnung von Pronomen. So hat beispielsweise die RE *he* den folgenden Wert für das Attribut HeSheItThey = „1,0,0,0". Dies bedeutet, dass mit einer Wahrscheinlichkeit von 100% mit dem Pronomen *he* auf die RE verwiesen wird. Die Wahrscheinlichkeit, dass mit dem Pronomen *she*, *it* oder *they* auf die RE *he* verwiesen wird, liegt bei 0%. Für den referentiellen Ausdruck *which* sieht der Wert des Attributes wie folgt aus: HeSheItThey=„0,0,0.5,0.5" [Postolache 2004:26]. Dies bedeutet, dass je mit einer Wahrscheinlichkeit von 50% auf den referentiellen Ausdruck *which* mit dem Pronomen *it* und dem Pronomen *they* verwiesen wird. Dieser Wahrscheinlichkeitswert wird mit Hilfe von WordNet-Einträgen errechnet.

Zuerst wird nach allen Synsets[57] des Hauptnomens der NP gesucht. Ihre Anzahl wird als **n** bezeichnet. Dann wird ermittelt, wie viele dieser Synsets Hypernyme der Synsets *female, female person* sind und ihre Anzahl wird als **f** vermerkt. Die Anzahl der Synsets, die Hypernyme der Synsets *male*, *male person* sind, wird als **m** notiert. Als **p** wird die Anzahl derjenigen Synsets bezeichnet, die Hypernyme der Synsets *person*, *individual*, *someone*, *somebody*, *mortal*, *human*, und *soul* sind. Die Anzahl der restlichen Synsets, die nicht Hypernyme der eben genannten Synsets sind, werden unter **t** aufgeführt. Die Wahrscheinlichkeit (P_{he}, P_{she}, P_{it} und P_{they}), mit welcher auf die RE mit den Pronomen *he*, *she*, *it* und *they* verwiesen wird, wird wie folgt berechnet. Wenn die RE im Plural steht, hat P_{they} den Wert 1 und P_{he}, P_{she}, und P_{it} den Wert 0. Andernfalls gilt „P_{he} = [m+1/2*(p-m)]/n, P_{she} = [f+1/2*(p-f)]/n, P_{it} = (n-p)/n und P_{they} = 0" [Postolache 2004:26].

Mit einer weiteren Wissensquelle können Antezedenten von *wh*-Pronomen bestimmt werden. Dazu werden Heuristiken verwendet. Eine solche Heuristik besagt, dass wenn eine nicht-verschachtelte NP einem *wh*-Pronomen vorangeht, sie als Antezedens des *wh*-Pronomens gilt (siehe Beispiel (78a)). Wenn die dem *wh*-Pronomen vorangehende NP verschachtelt ist, gilt folgendes: Die erste eingebettete definite NP wird als Antezedens angesehen (78b). Falls es eine solche NP nicht gibt, wird die äussere NP als Antezedens genommen (78c).

[57] Die Information in WordNet ist in Synsets organisiert. Synsets sind logische Gruppierungen von Wörtern. Sie enthalten synonyme Wörter und Kollokationen.

(78) a *He sat down at **a small table** that stood to the left of the telescreen.*
b *They were [the homes of **[the four Ministries]**] between which the entire apparatus of government was divided.*
c *He remained within **[the field of [vision]]** which the metal plaque commanded.*
[Postolache 2004:26]

Weiter wurden auch anaphorische Relationen wie "belongs-to, type-of, is-part-of, has-as-part, is-element-of, has-as-element, same-as" [Cristea et al. 2005:39] manuell ermittelt. Mit diesen Informationen konnten auch Anaphern komplexerer Art aufgelöst werden.

Eine weitere Komponente des Systems sind Matching-Regeln. Mit der Regel namens IncludingRule können unpassende Antezedenskandidaten ausgeschlossen werden. Die Regel nimmt eine PS und eine DE als Argumente. Wenn die RE, anhand der die PS generiert wurde, in irgendeiner RE eingebettet ist, die auf die DE verweist, kann die PS nicht auf diese Einheit verweisen. Neben ausschliessenden Regeln, gibt es auch Regeln die einen möglichen Kandidaten als Antezedens bestätigen. Die ProperNameRule besagt, dass eine PS, die einen Personennamen bezeichnet, sich auf eine DE bezieht, auf die bereits mit demselben Personennamen verwiesen wurde. Diese Regel verwendet die Merkmale femaleName, maleName, familyName und lemma. Die Regel WhRule, die vom wh Attribut Gebrauch macht, bestätigt, dass eine PS und die dazugehörende RE auf eine DE verweist, wenn die Identifikationsnummer (ID) des Antezedens der PS der ID einer der REs entspricht, die die besagte DE konstituieren. Die AppositionsRule attestiert, dass die Apposition einer RE mit der RE koreferent ist.

Zudem gibt es auch Regeln, anhand denen die für eine PS in Frage kommenden DEs bewertet werden. Die NumberRule retourniert den Wert 1, wenn die PS und die DE denselben Wert für das Attribut number haben, und 0 wenn dies nicht der Fall ist. Es ist anzumerken, dass eine DE auch verschiedene Werte für ein und dasselbe Attribut haben kann, da die Werte von allen auf diese DE verweisenden REs zusammengefasst sind. So wird mit den Regeln eigentlich geprüft, ob der Wert eines PS-Attributs mit einem Wert des entsprechenden DE-Attributs übereinstimmt. Die Regel LemmaRule, die das Attribut lemma verwendet, ist der NumberRule ähnlich. Die Regel RoleRule gibt Werte zwischen 0 und 1 aus, wobei Subjekte höher gewertet werden als direkte Objeke, direkte Objekte höher als indirekte Objekte, und indirekte Objekte höher als Attribute. Mit der Regel SynonymyRule[58] wird in WordNet nach einem Synset gesucht, das sowohl das Lemma der PS als auch der DE beinhaltet. Wird ein solches Synset gefunden, wird der Wert 1 ausgegeben, andernfalls der Wert 0. Mit der Regel WordNetChain-

[58] Es kann natürlich sein, dass die Lemmata in einem Text unterschiedliche Bedeutungen haben wie im Beispiel „Er setzte sich auf die Bank vor der Kreditanstalt." In diesem Fall verweist die RE Kreditanstalt nicht auf die gleiche DE wie die RE Bank. Die SynonymyRule würde hier also eine inkorrekte Verbindung andeuten. Daher erhält diese Regel von [Postolache 2004:28] eine relativ tiefe Gewichtung in der Gesamtwertung.

Rule[59] wird in WordNet nach Hyperonymie- oder Hypernymie-Relationen zwischen dem Lemma der PS und dem Lemma der DE gesucht. Kann eine solche Relation festgestellt werden, erhält man einen Wert 1, andernfalls einen Wert 0.

Nachdem die Komponenten beschrieben worden sind, wird nun auf die Verarbeitung der Anaphern im RARE-System eingegangen. Entweder wird eine RE klar einer DE zugeordnet oder ein Matching-Wert für die aktuelle RE anhand der PS und den in Frage kommenden DEs errechnet. Wenn dieser Wert unter einem bestimmten Minimum liegt, wird eine neue DE für die aktuelle RE anhand der PS generiert, da keine vorangehende DE für die neue RE in Frage kommt. Die PS wird danach gelöscht und, wie oben erwähnt, eine direkte Verbindung von der RE zu der DE hergestellt.

Wenn es DEs gibt, die einen Wert über dem bestimmten Minimum erzielen, wird die DE mit dem besten Wert als Antezedens der aktuellen RE angesehen, falls keine weitere DE innerhalb eines bestimmten Differenzbereichs auftritt. Auch in diesem Fall wird die zur RE gehörende PS gelöscht. Falls eine weitere DE innerhalb des bestimmten Differenzbereichs auftritt, bedeutet dies, dass sich die DE mit dem besten Wert zu wenig von anderen DEs absondert und die RE vorläufig keine DE zugeordnet bekommt. Die Zuordnung eines Antezedens wird also verschoben und die PS bleibt vorübergehend bestehen. Aufgelöste PSs fügen ihre Attribute den ihnen zugeordneten DEs zu. So kann anhand der neu gewonnenen Information eine RE zu einem späteren Zeitpunkt aufgelöst werden.

Wie [Cristea et al. 2002:1] schreibt, spiegelt das System den Leseprozess wider: Die meisten Anaphern können sofort bei ihrem Erscheinen aufgelöst werden. In manchen Fällen muss die Interpretation der Anapher aber verschoben werden, bis zusätzliche Information zur Auflösung beitragen kann. Im Beispiel (79) muss die Auflösung des Pronomens *she* verschoben werden bis der zweite Satz vollständig interpretiert werden kann. Dank einer semantischen Einschränkung, die besagt, dass wohl dieselbe Person, die lernt, ihre Noten verbessern möchte, können die Pronomen *her* und *she* Emily zugeordnet werden.

(79) *When Roberta came home Emily was studying.* ***She*** *really wanted to improve* ***her*** *grades.*

[Postolache 2004:21]

Am Ende des Prozesses sollte jede RE eine DE zugewiesen haben. Wie [Postolache 2004:29] schreibt, konnten mit dem System Erfolgsraten zwischen 60% und 70% erreicht werden.

[59] Wie die SynonymyRule erhält auch die WordNetChainRule eine relativ tiefe Gewichtung in der Gesamtbewertung. Die Begründung ist dieselbe.

Nachdem im Kapitel 4 die Schwierigkeiten mit den verschiedenen Typen von Anaphern beleuchtet worden sind, wird im Kapitel 5 die Problematik der Diskrepanzen in den Kongruenzmerkmalen zwischen der Anapher und ihrem Antezedens aufgegriffen.

5 Diskrepanzen in den Kongruenzmerkmalen

Der Prozess der AR kann in drei Phasen gegliedert werden: Die Identifikation der Anapher, das Finden von möglichen Antezedenten und die Auswahl eines Antezedens. Die Auswahl des Antezedens hängt von bestimmten Faktoren ab. Diese Faktoren werden in Einschränkungen und Präferenzen unterteilt und im Folgenden diskutiert.

Neben den Schwierigkeiten, welche die einzelnen Faktoren betreffen, gibt es auch allgemeine Probleme betreffend der Faktorenwahl, der Reihenfolge sowie auch der Auflösungsstrategie. Es muss entschieden werden, welche Faktoren als Einschränkungen dienen, welche Präferenzen sind und ob bestimmte Faktoren überhaupt als einschränkend angesehen werden sollen. Wie [Mitkov 1997:20] schreibt, sind Präferenzen sicherer, da sie nicht ausschliessend sind; Einschränkungen können aber die Auswahlprozedur von Antezedenskandidaten beschleunigen.

[Mitkov 1997:15] hat zwei Auflösungsstrategien miteinander verglichen und festgestellt, dass eine kombinierte Strategie wohl die beste Lösung ist. In den beiden von [Mitkov 1997:15] untersuchten Systemen wurden die gleichen Faktoren verwendet. Während in einem System bestimmte Faktoren als Einschränkungen dienten und andere als Präferenzen, gab es im anderen System nur Präferenzen, denen Werte zugeordnet wurden. Wie die Untersuchung von [Mitkov 1997:19] ergab, erzielten beide Ansätze ähnliche Resultate. Erfolgreicher als die beiden Ansätze war allerdings eine Hybrid-Methode, die beide Ansätze beinhaltete. Der Vorteil dieser Methode war gemäss [Mitkov 1997:18], dass der Suchprozess, falls die Ergebnisse der beiden Ansätze übereinstimmten, früher abgebrochen werden konnte als in den einzelnen Systemen.

Weiter muss in Bezug auf die Faktoren beachtet werden, dass es auch genreabhängige Faktoren gibt. Laut [Mitkov 1997:20] scheinen die Faktoren „Objektpräferenz“ und „Titelpräferenz“ genrespezifisch zu sein. Ausserdem können Faktoren voneinander abhängig sein, wie beispielsweise die Objektpräferenz, die durch Verben angezeigt wird, und die Objektpräferenz, die durch Nomen angedeutet wird [Mitkov 1997:19]. Bei einem Gewichtungsmechanismus kann es daher von Vorteil sein, nur einen dieser Faktoren zu benützen. Solche Abhängigkeiten zwischen Faktoren müssen allerdings detaillierter untersucht werden, um sie im Auflösungsprozess bestmöglich berücksichtigen zu können.

Zur Auflösung von Anaphern werden syntaktische und semantische Einschränkungen verwendet. Gemäss der *Government and Binding*-Theorie[60] kann die syntaktische Struktur

[60] Die GB ist eine Version von Noam Chomskys Syntaxmodell der generativen Transformationsgrammatik

Anforderungen an intrasententielle Antezedenten stellen. Das in der Theorie verwendete Konzept heisst *C-Command*. Ein *C-Command*-Knoten dominiert in einem Parsebaum seine Geschwisterknoten und alle Knoten, die von den Geschwisterknoten dominiert werden. Die drei wichtigsten Regeln lauten, dass eine nicht-pronominale NP nicht auf eine sie c-dominierende NP verweisen kann (80a), dass das Antezedens einer gebundenen Anapher die Anaphern c-dominieren muss (80b), und dass ein Personalpronomen sich nicht auf eine es c-dominierende NP beziehen kann (80c). Vergleiche hierzu Abbildung 44.

(80) a *He$_i$ likes Pat$_j$.*
b *Pat$_i$ likes himself$_i$.*
c *Pat$_i$ likes him$_j$.*

[Byron 2004a:33,34)

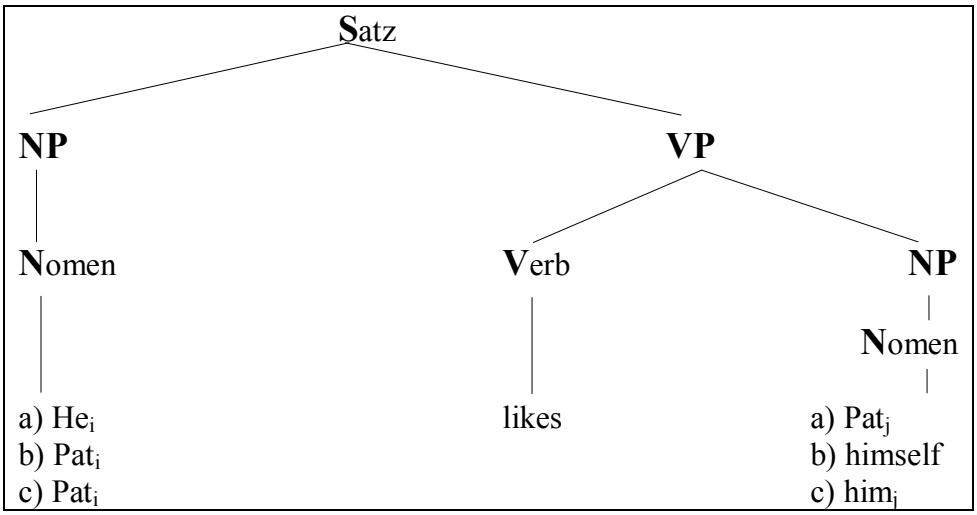

Abb. 44 Syntaxbaum für die Sätze in Beispiel (80)

[Byron et al. 2004b:21] haben versucht syntaktische Merkmale zu benutzen, um indefinite NPs zu identifizieren, die in keiner anaphorischen Beziehung stehen und deshalb nicht als Antezedenten von Pronomen in Frage kommen. In dem Satz „*Pat is a great teacher*“ [Byron et al. 2004b:21] sollte für ein nachfolgendes Pronomen nur *Pat* als Antezedens zugänglich sein, nicht aber die Phrase *a great teacher*. Solche indefiniten NPs kommen auch in Konstruktionen mit Negationen (81a), Modalverben (81b) und negativen Implikationen (81c) vor. Weiter können sie auch in Sätzen mit Verben wie *try, intend, plan* auftauchen, die eine noch unwahre Proposition beinhalten und kommen auch in Befehlen und Ja- und Nein-Fragen vor.

(81) a *Bill does not have a car.*
b *Bill can make a kite.*
c *John forgot to write a term paper.*

[Byron et al. 2004b:22]

Abbildung 45 zeigt, wie indefinite NPs, die nicht als Antezedens zugänglich sind, ausgefiltert werden können.

```
NON-REF if (appositive is true
            or determiner is 'no'
            or numeric value is true
            or (the NP is indefinite
            and (negated is true
                 or predicate is true
                 or modal is true
                 or modifier is true))
```

Abb. 45 Code zur Ausfilterung von indefiniten NPs, die nicht als Antezedenten zugänglich sind [Byron et al. 2004b:22]

Eine Einheit wird nicht als Antezedens für ein anaphorisches Pronomen gewählt, falls sie in einer Apposition steht, keinen Artikel hat oder ein numerischer Wert ist. Indefinite NPs können auch nicht als Antezedenten fungieren, wenn sie in einer Verneinung oder Konstruktionen mit Kopula stehen, durch ein Modalverb dominiert werden oder Teil einer Modifikation sind. Die Ergebnisse von [Byron et al. 2004b:21] haben gezeigt, dass die eben erwähnte Ausfilterung von Antezedenskandidaten nur eine sehr geringe Verbesserung der Pronomenresolutionsalgorithmen, die auf Syntaxstrukturen basieren, ergibt. Techniken, die Syntaxstrukturen nicht einbeziehen, können allerdings von dem Filter profitieren.

[Byron 1999:42] schlägt vor, das Tempus in die AR von Pronomen einzubeziehen. Einheiten, die in der Vergangenheit erwähnt werden, können nicht in einem anderen Zeitrahmen aufgegriffen werden. (z.B. „*I'm going to bake some muffins tomorrow. *They tasted great*.").

Zu den semantischen Einschränkungen gehören Numerus-, Genus- und Belebtheitskongruenz. Es können aber auch durch Verben und Adverbien Anforderungen an das Antezedens gestellt werden. Im Beispiel (82) kann sich das Pronomen *it* nur auf den Kuchen beziehen, da das Verb *eat* ein essbares Objekt verlangt.

(82) *Pat took **the cake** from the table and ate **it**.*

[Byron 2004a:35]

Natürlich kann, wie bereits erwähnt, auch Weltwissen und Kontextwissen zur Auflösung von Anaphern beitragen, jedoch ist die Implementierung von solchem Wissen sehr schwierig.

Im Folgenden soll auf Schwierigkeiten mit den wichtigsten Kongruenzmerkmalen (Numerus, Genus und Belebtheit) eingegangen werden.

5.1 *Diskrepanzen im Numerus*

Die Numeruskongruenz dient in vielen AR-Systemen als wichtiges Einschränkungskriterium und trägt zur Auswahl des korrekten Antezedens bei. Oft kann dies aber zu einer Fehlinterpretation von Anaphern führen, da Numerusinkongruenz ein häufiges Phänomen in der natürlichen Sprache ist.

Syntaktisch gesehen darf der Numerus der Anapher vom Numerus des Antezedens abweichen. Zwar muss der Numerus der Anapher mit dem Numerus der Diskurseinheit (der semantischen Realisation des Antezedens) übereinstimmen, jedoch können sich die semantische und die syntaktische Form des Antezedens in Bezug auf den Numerus unterscheiden. Obwohl im Beispiel (83) die Nominalphrase *each child* syntaktisch gesehen eine Singularform ist, sind semantisch betrachtet alle Kinder der Klasse gemeint, weshalb mit der Pluralanapher *they* auf das Antezedens verwiesen wird. Die Anapher stimmt also in Bezug auf den Numerus mit der Diskurseinheit und nicht mit deren syntaktischen Ausführung überein.

(83) *The teacher gave* ***each child*** *a crayon.* ***They*** *started drawing colorful pictures.*
[Mitkov 2002:8]

Im umgekehrten Fall, kann auch eine Anapher im Singular ein Antezedens im Plural haben, wie aus den Beispielsätzen in (84) ersichtlich wird.

(84) *I met* ***two doctors*** *yesterday.* ***The tall one*** *told me a story.*
[Clark 1977:171]

Weiter muss beachtet werden, dass bestimmte Nomen wie *fish* in Bezug auf ihren Numerus ambig sind.

[Barbu et al. 2002:1,3] haben experimentell bestätigt, dass beinahe ein Viertel von 2500 von ihnen untersuchten Pluralpronomen aufgrund von Numerusinkongruenz nicht dem korrekten Antezedens zugeordnet werden kann. [Barbu et al. 2002:2,3] haben diese Unstimmigkeiten in acht Gruppen geteilt, die in Abbildung 46 aufgezeigt werden.

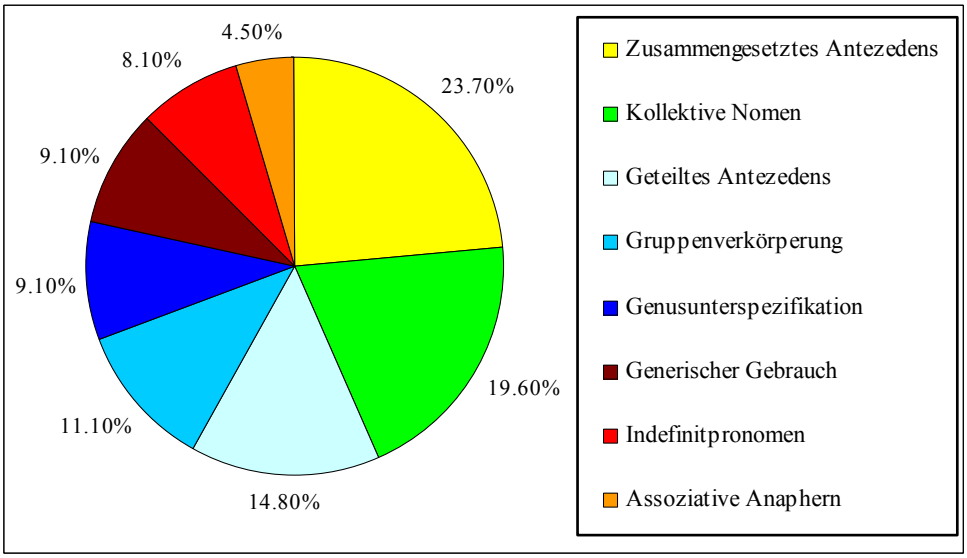

Abb. 46 Numerusinkongruenz zwischen Pluralanpahern und ihrem Antezedens, basierend auf [Barbu et al. 2002:3]

Es soll nun näher auf die syntaktische Numerusinkongruenz zwischen Anaphern und ihren Antezedenten eingegangen werden.

5.1.1 Numerusdiskrepanzen aufgrund komplexer Antezedenten

Inkongruenz im Numerus geht oft mit komplexen Antezedenten einher. Wie aus den Beispielsätzen in (85) hervorgeht, können NPs durch eine Konjunktion oder Disjunktion[61] zu einem zusammengesetzten Antezedens verbunden werden. Das Antezedens kann sich aber auch aus mehreren auseinander gerissenen NPs zusammensetzen. Im Gegensatz zu [Barbu et al. 2002:3] wird hier nicht zwischen zusammengesetzten und geteilten Antezedenten unterschieden. Diese werden als komplexe Antezedenten zusammengefasst.

(85) a *Michael **and** Maria were at the cinema. They had a great time.*
b *Michael was at the cinema **with** Maria. They had a great time.*
c *Michael **met** Maria at the cinema. They had a great time.* [Eschenbach et al. 1989:161]
d *Joe **took** the boys to the ball game. They had a good time.* [Denber 1998:16]

[61] Beispiel: "Working in such a school in the mid-1950s, part of my task as a very inexperienced junior teacher was to explain almost daily to parents that **Peter or John or Mary** would learn their tables, and learn to compute, despite the fact that **they** were using apparatus rather than sum books and were active rather than passive." [BNC: CLY 1677]

Neben Konjunktionen wie *and* und *or* können auch einige Präpositionen wie *with* und bestimmte Verben wie *meet* und *take* auf komplexe Antezedenten hindeuten. PPs weisen allerdings nicht notwendigerweise auf zusammengesetzte Antezedenten hin. Während sich im Satz (86a) die Pluralanapher *they* auf das komplexe Antezedens bestehend aus Michael, Maria und Peter beziehen kann, bezieht sie sich im Satz (86b) nur auf Michael und Maria.

(86) a *Michael and Maria were at the park with Peter. In the evening they were at a garden party.*
b *Michael and Maria were at the park with their Frisbee. In the evening they were at a garden party.*
[Eschenbach et al. 1989:161]

Wie kann nun automatisch erkannt werden, wann eine Einheit zu einem komplexen Antezedens gehört?
Als Lösung bietet [Eschenbach 2000:162] das Konzept der „Common Association Basis (CAB)“ an, das beinhaltet, dass ein komplexes Antezedens nur aus Einheiten des gleichen ontologischen Typs zusammengesetzt sein darf. In Beispiel (86b) wird *Frisbee* nicht zur zusammengesetzten Einheit Michael und Maria gezählt, da Dinge und Menschen nicht vom selben ontologischen Typ sind.
Bei den Verben muss laut [Eschenbach 2000:163] zwischen verbindenden und trennenden Verbkonzepten unterschieden werden. Im Gegensatz zum Verb *meet*, muss das Verb *watch* eher als trennendes Verbkonzept betrachtet werden. Im Beispiel (87a) wird das komplexe Antezedens bestehend aus Michael, Peter und Maria dem bestehend aus Peter und Maria vorgezogen. Im Beispiel (87b) tritt der umgekehrte Fall ein. Es wird das Antezedens bevorzugt, das aus Peter und Maria zusammengesetzt wird.

(87) a *Michael met Peter and Maria in the pub. They had a great time.*
b *Michael watched Peter and Maria in the pub. They had a great time.*
[Eschenbach et al. 1989:163]

In Beispiel (88) kann aber nicht eindeutig bestimmt werden, ob sich das Pronomen *they* auf das komplexe Antezedens Michael und Maria oder auf das komplexe Antezedens Peter und Anne bezieht. Für die korrekte Auflösung ist Kontextwissen notwendig.

(88) *Michael and Maria picked up Peter and Anne from the station. They were late.*
[Eschenbach et al. 1989:165]

Auf die Implementation seines Ansatzes geht [Eschenbach et al. 1989] nicht genauer ein. Die Schwierigkeit ist, die erforderliche Information aus Ontologien zu gewinnen und sinnvolle Listen mit den verschiedenen Verbkonzepten zu erstellen. Für ambige Fälle ist zudem Kontextwissen erforderlich, das kaum implementiert werden kann.

[Cristea et al. 2005:33] sind ebenfalls der Meinung, dass Numerusinkongruenz mit Ähnlichkeitskriterien[62] angegangen werden sollen. Wenn auf *Athos*, *Porthos* und *Aramis* mit der definiten NP *the musketeers* verwiesen wird, kann eine komplexe Diskurseinheit [*Athos, Porthos, Aramis*] aufgrund der Information gebildet werden, dass die Namen auf Menschen verweisen und dass Musketiere ebenfalls Menschen sind. Diese Information muss der Maschine allerdings anhand von Ontologien oder Listen zur Verfügung stehen.
Im Gegensatz zu [Eschenbach 1998] sind [Cristea et al. 2005:33] aber der Meinung, dass komplexe Antezedenskandidaten, wie [*Michael, Maria*] und [*Michael, Maria, Peter*] in Beispiel (86) nicht automatisch beim Parsen von *Maria* bzw. *Peter* zur Verfügung gestellt werden müssen, sondern erst mit dem Auftauchen der Pluralanapher *they*. In ihrem Ansatz [Cristea et al. 2005:33] wird eine neue, komplexe Diskurseinheit vorgeschlagen, wenn kein geeignetes Antezedens für die Pluralanapher gefunden werden kann, d.h., wenn kein Pluralantezedens im Suchbereich zu finden ist. Ist dies jedoch der Fall, scheint der Pluralkandidat aber ungeeignet, wird die komplexe Diskurseinheit zwar gebildet, die Auflösung aber verschoben.
Die Auflösung von Pluralpronomen, die sich auf ein komplexes Antezedens beziehen, kann laut [Barbu et al. 2002:4] mit einer kleinen Anzahl an Regeln gemeistert werden. Eine zusammengesetzte NP sollte als Antezedens für eine Pluralanapher bereitstehen, ihren Konstituenten sollte dies aber untersagt sein. So sollte es im Satz (89) für das Pronomen *they* nicht möglich sein, sich nur auf Tom und John zu beziehen.

(89) ***Tom**, **John** and **Mary** went to the cinema. **They** saw a comedy.*
[Barbu et al. 2002:4]

5.1.2 Numerusdiskrepanzen aufgrund kollektiver Nomen

Grammatikalische Numerusdiskrepanzen treten häufig auch in Zusammenhang mit kollektiven Nomen wie *the police* auf. Kollektive Nomen beinhalten auch Firmennamen und Organisationsnamen, auf die oft mit einer Singular- oder Pluralanapher verwiesen werden kann.

[62] in [Eschenbach 1998] CAB genannt

Wie [Barbu et al. 2002:5] schreiben, können kollektive Nomen anhand eines Lexikons ausgefiltert werden. Auch NPs, die durch ein Kollektivnomen wie a *number of* und a *group of* modifiziert werden, können markiert werden. Firmennamen müssen allerdings mit einem Namenserkennungssystem erkannt werden, oder es müssen zumindest Regeln eingebaut werden, die ein Nomen in Grossbuchstaben gefolgt von Abkürzungen wie beispielsweise *co.* und *ltd.* und Nomen wie *Corporation* und *Incorporated* als Firmennamen ansehen. In [Mitkov 2002a:146] werden kollektive Nomen insofern berücksichtigt, dass ihnen die Numerus-Prüfung erlassen wird.
Weiter gibt es NPs, die syntaktisch gesehen im Plural stehen, aber semantisch wie Einzahlformen agieren, wie beispielsweise die NP *the movies* (= das Kino).

5.1.3 Numerusdiskrepanzen aufgrund von Gruppenverkörperungen

Auch Gruppenverkörperungen können zu Numerusinkongruenz führen, wie aus dem Beispiel (90) ersichtlich ist. Eine NP im Singular kann eine Gruppe von Einheiten verkörpern und somit besteht die Möglichkeit auf diese NP mit einer Pluralanapher zu verweisen.

(90) *I need **a plate**. Where do you keep **them**?*

[Barbu et al. 2002:2]

5.1.4 Numerusdiskrepanzen aufgrund von Genusunterspezifikation

Numerusinkongruenz kommt auch in Zusammenhang mit Genusunterspezifikation vor, wie Beispiel (91) zeigt. Nicht-genusmarkierte Pronomen werden dabei verwendet, um die Phrase *he or she* zu ersetzen.

(91) *Once **a person** registers for an interview **they** are removed from the unemployment figures.*

[Barlow 1998:35]

Dies ist auch bei der Verwendung von Indefinitpronomen wie beispielsweise *everyone*, *nobody* und *somebody* der Fall.

(92) ***Someone** will remember to wake me up early in the morning, won't **they**?*

[Barbu et al. 2002:3]

Da auf Indefinitpronomen meistens mit Pluralpronomen verwiesen wird, können auch hier laut [Barbu et al. 2002:5] einfache Regeln eingebaut werden. Die Auflösung von Pluralpronomen, die auf eine NP im Singular verweisen, ist hingegen weit komplexer. Man kann zwar indefiniten NPs erlauben, als Antezedens einer Pluralanapher in Frage zu kommen, dies kann allerdings zu weiteren Fehlern führen, da die Anzahl der möglichen Kandidaten erhöht wird.

5.1.5 Numerusdiskrepanzen aufgrund assoziativer Anaphern

Auch in Zusammenhang mit assoziativen Anaphern ist Numerusinkongruenz anzutreffen. Oft kann nur eine Person explizit im Text erwähnt sein, aus dem Kontext lässt sich aber schliessen, dass mehrere Personen involviert sind. So signalisiert die Hochzeit im Beispiel (93), dass es neben der Braut auch einen Bräutigam gibt.

(93) *My sister's wedding was beautiful.* ***They*** *were the happiest couple.*
[Barbu et al. 2002:3]

NP-Anaphern können sich sowohl im Lemma als auch im Numerus von ihrem Bezugswort unterscheiden, wie anhand des Beispiels (94) aufzeigt wird.

(94) *a patrol… the soldiers*
the government... the ministers
[Postolache 2004:8]

[Postolache 2004:9] schlägt vor, in solchen Fällen in einer Ontologie nachzuschlagen und wenn eine der Bedeutungen des Nomens auch eine Gruppe beinhaltet, einen entsprechenden Eintrag im System zu machen. Da eine der Bedeutungen von *patrol* und *government* auch das Konzept der Gruppe beinhaltet, könnten die Pluralanaphern *the soldiers* und *the ministers* richtig aufgelöst werden. Allerdings kann es laut [Postolache 2004:9] bei diesem Ansatz zu Fehlern kommen, da viele Wörter eine Bedeutung haben, die mit dem Konzept der Gruppe einhergeht. Zum Beispiel ist eine der Bedeutungen des Wortes *garden* als Gruppe von Blumen definiert, was in einem Satz wie (95) zur falschen Auflösung der Anapher *the soldiers* führen kann.

(95) *A patrol stopped by the garden.* ***The soldiers****...*
[Postolache 2004:9]

Deshalb müssen laut [Postolache 2004:9] mehrere Kriterien kombiniert werden wie beispielsweise Ähnlichkeitskriterien oder allenfalls eine Liste, die kollektive Nomen enthält.

5.2 *Diskrepanzen im Genus*

Eine Anapher muss mit ihrem Antezedens im Genus übereinstimmen. Da sich das Englische nicht durch grammatikalisches Genus auszeichnet, ist die Genuskongruenz im Englischen von geringerer Bedeutung als in anderen Sprachen, wie beispielsweise dem Deutschen. Das Problem der Genuskongruenz kann im Englischen somit auf einige wenige Anomalien, auf Eigennamen und auf Übereinstimmung in der Belebtheitsinformation reduziert werden. Im folgenden Kapitel wird darauf näher eingegangen.

5.2.1 Genusdiskrepanzen aufgrund von Anomalien

Anomalien in Bezug auf das Genus sind im Englischen eher selten. So kann beispielsweise auf Schiffe mit dem weiblichen Pronomen *she* verwiesen werden, obwohl es sich um unbelebte Objekte handelt. Ausserdem können Naturerscheinungen und Prozesse vermenschlicht und somit mit maskulinen oder femininen Pronomen wieder aufgegriffen werden, wie Beispiel (96) verdeutlicht.

(96) *Ol' man River, he jus' keep rollin'.*
When the print manager sees a new job, he queues it up.
[Denber1998:18]

5.2.2 Genusdiskrepanzen aufgrund von Eigennamen

Das Problem mit Eigennamen ist einerseits, dass sie oft von anderen Einheiten abgeleitet werden und daher von diesen unterschieden werden müssen. Andererseits sind Eigennamen in Bezug auf ihr Geschlecht nicht immer eindeutig.

[Denber 1998:17] hat einen Ansatz zur Erkennung von Eigennamen entwickelt. Sein Algorithmus beruht auf der Tatsache, dass Eigennamen im Englischen grossgeschrieben sind. Allerdings beginnen auch Wörter am Satzanfang und Titel (*Count Dracula*, *Earl of Surrey*, *Lady Diana*) mit einem Grossbuchstaben. Ausserdem können Texte vollumfänglich in Grossbuchstaben verfasst sein. Dies ist insbesondere problematisch, da Namen auch gewöhnliche Nomen sein können. So wird beispielsweise ein Wagenheber im Englischen auch „*Jack*" [Denber 1998:17] genannt und Mädchennamen sind im Englischen oft von Blumennamen wie

„*Daisy, Heather, Ivy, Rose*" oder Monatsnamen wie „*April*" [Denber 1998:17] abgeleitet. Auch können Personennamen mit Ortsnamen identisch sein, wie beispielsweise „*Troy*" [Denber 1998:17]. Schliesslich gibt es auch besondere Namen wie *Cosma Shiva* oder *Moon Unit*, Übernamen wie *Chalky* und *Sandy*, Künstlernamen wie *Nicolas Cage*, *John Cougar* und *Alice Cooper* und Bandnamen wie *The Doors* und *The Rolling Stones.*

Problematisch ist auch, dass einige Eigennamen in Bezug auf ihr Genus ambig sein können, wie „*Chris, Pat*" und „*Tracy*" [Denber 1998:17] oder sich das Genus von Namen in verschiedenen Sprachen unterscheiden kann. So ist, wie [Denber 1998:17] schreibt, „*Jean*" ein Mädchenname im Englischen und ein Knabenname im Französischen.

Eine weitere Schwierigkeit in Zusammenhang mit Namen sind Vornamen gefolgt von Nachnamen sowie mehrfache Vornamen. Da sich die Anapher auf den gesamten Namen bezieht, muss das System diverse Namensteile als Einheit interpretieren können. In Beispiel (97) verweist das Pronomen *he* auf beide Vornamen und den Nachnamen, also auf *Edgar Allan Poe*.

(97) *Everybody knows* ***Edgar Allan Poe***. ***He*** *is considered one of the most famous writers of American literature.*

[Denber 1998:18] baut als Abhilfe eine Heuristik ein, die das Wort nach dem Eigennamen prüft. Wenn es sich ebenfalls um einen Eigennamen handelt oder um ein Nomen unbestimmten Typs wird es mit dem vorangehenden Namen zu einem Kompositum verbunden.

Genus- und Belebtheitsinformationen lassen sich anhand von Schlüsselwörtern gewinnen. Anreden wie Mrs. und Ms. deuten darauf hin, dass es sich um eine weibliche Person handelt und Abkürzungen wie Inc. und Corp. werden mit Firmen assoziiert. Nomen wie father, daughter, woman und widow signalisieren Genusinformation, die laut [Siddharthan 2003:9] aus WordNet gewonnen wird. Weiter kann geprüft werden, ob eine NP nicht mit einer NP koreferiert, die bereits die nötigen Merkmalswerte enthält (z.B. „Pierre Vinken ... Mr. Vinken" [Siddharthan 2003:9]). Auch aus Appositionen und Kopula-Konstrukten lassen sich manchmal Genus- und Belebtheitsinformationen ableiten, wie Beispiel (98) verdeutlicht.

(98) ***Finmeccanica*** *is* ***an Italian state-owned holding company*** *with interests in the mechanical engineering industry.*

[Siddharthan 2003:10]

5.3 *Belebtheitsinformation zur Auflösung von Anaphern*

Für die AR ist auch Belebtheitsinformation von Bedeutung. Um die Anapher *he* in dem Beispiel (99) korrekt aufzulösen, muss das System ein belebtes Antezedens finden, beziehungsweise *a dog* als belebte Einheit erkennen.

(99) *A picture of a dog carrying a stick. He is running in circles.*
[Denber 1998:19]

Belebtheitsinformation lässt sich ebenfalls aus WordNet ermitteln. WordNet ist dafür sehr geeignet, da Belebtheitsinformation nicht aus textuellen Merkmalen gewonnen werden kann, sondern vielmehr im Weltwissen verankert ist. Ein weiterer Vorteil ist, dass auch Verbinformation beigezogen werden kann. Der Ansatz von [Denber 1998:19] sucht in der WordNet-Hierarchie des zu klassifizierenden Nomens nach dem Schlüsselwort „*creature*". Falls eine solche Information gefunden wird, wird das Nomen als belebt markiert. [Siddharthan 2003:9] kennzeichnet ein Nomen als belebt, wenn es ein Hyponym der WordNet Kategorien *human*, *animal* oder *organisation* ist. [Evans et al. 2000: 158,159] geht davon aus, dass Hierarchien mit den Konzepten *animal*, *person* und *relation* belebte Einheiten umfassen. Auch Verben können Belebtheitsinformationen implizieren. So haben normalerweise Verben wie *said, report, state, ask* belebte Subjekte. Laut [Evans et al. 2000:159] sind es die Hierarchien *cognition*, *communication*, *emotion* und *social*, die als Subjekt eine belebte NP vermuten lassen. WordNet kann also zur Erschliessung von Belebtheitsinformationen verwendet werden.
Problematisch ist wiederum die Behandlung von Eigennamen. Wie [Denber 1998:19] schreibt, ist die Einheit *George Washington* belebt, die NP *the George Washington Bridge* aber nicht. Das System muss natürlich *George Washington Bridge* als eine Einheit erkennen können. Der Algorithmus von [Denber 1998:19] schliesst solche Fälle aber nicht mit ein.
Die Zuteilung von Belebtheitsinformationen anstelle von spezifischen Genusinformationen kann in der englischen Sprache von Vorteil sein. Zum einen gibt es im Englischen nebst den Pronomen der 3. Person Singular nur wenige sprachliche Hinweise, die auf das Genus einer NP hindeuten. So sind dies einige Suffixe, wie *–ess*, *–man* oder -*woman* die auf das Geschlecht hinweisen, wie beispielsweise in den Wörtern *princess*, *fireman,* und *chairwoman*. Jedoch wird das Suffix *–ess* in der Gegenwart immer weniger verwendet und das Suffix *–man* kann auch gebraucht werden, wenn von weiblichen Personen die Rede ist.
Zum anderen muss man in bestimmten Fällen einer NP mehrere Genera zuordnen, damit die Übereinstimmung zwischen dem Pronomen und dem Antezedenten nicht beeinträchtigt wird. Da es sich im Beispiel (100) um eine Beschreibung von Leuten ungeachtet ihres Geschlechts

handelt, muss die NP *the user* sowohl mit maskuliner als auch femininer Genusinformation versehen sein.

(100) ***the user of the machine*** *should select* ***his*** *or* ***her*** *own settings*
[Evans et al. 2000:156]

Ausserdem ist die automatische Zuordnung des Genus weitaus problematischer als die automatische Zuteilung von Belebtheitsinformationen. Genusinformation kann im Englischen nur bei der Auflösung von Personalpronomen, Possessivpronomen und Reflexivpronomen im Singular helfen.
Obwohl Belebtheitsinformation zum Erfolg eines AR-Systems beitragen kann, ist sie in gewissen Genres (z.B. Wissenschaft, Technik) unnütz. Im Satz (101) sind *cashiers* keine Personen, sondern Teile des Computersystems.

(101) *when the cashiers need to get their sub-totals, they can exchange information…*
[Evans et al. 2000:156]

Wenn in einem Text das Vorkommen von belebten Einheiten dasjenige von unbelebten Einheiten bei weitem übersteigt, ist die genauere Unterscheidung zwischen maskulin, feminin und neutrum sinnvoller. Wie kann Belebtheitsinformation aber automatisch ermittelt werden?

5.3.1 Algorithmus zur Bestimmung der Belebtheitsinformation

[Evans et al. 2000:158-161] haben einen Algorithmus zur Bestimmung von Belebtheitsinformationen entwickelt. In einem ersten Schritt wird das Lemma des Hauptnomens einer NP bestimmt. Befindet sich die NP in einer Subjektposition, so wird auch das Lemma des Verbs, das die NP zum Subjekt hat, bestimmt. In einem zweiten Schritt wird WordNet nach dem Lemma des Hauptnomens der NPs durchsucht. Es wird die Anzahl der Bedeutungen des Nomens ermittelt, die zu einer der drei Hierarchien *animal*, *person* und *relation* gehören. [Evans et al. 2000:159] spricht hier von „Noun Animacy"-Bedeutungen (#*NAS*). Weiter wird auch die Anzahl der Bedeutungen des Nomens gezählt, die zu anderen Hierarchien gehören, von [Evans et al. 2000:159] „Noun Non Animacy"-Bedeutungen (#*NNAS*) genannt. Im Falle von NPs in der Subjektposition werden die Zählungen in Bezug auf das Verb, das die NP als Subjekt nimmt, durchgeführt. Es wird die Anzahl der Verbbedeutungen, die zu den Hierarchien *cognition*, *communication*, *emotion* und *social* gehören und somit ein belebtes Subjekt vermuten lassen (#*VAS*), der Anzahl der Bedeutungen, die zu anderen Hierarchien gehören

(#*VNAS*), gegenübergestellt. Anhand der Formeln in Abbildung 46 berechnet [Evans et al. 2000:156] Kennzahlen für die Belebtheit (Noun Animacy*; NA*) sowie für die Unbelebtheit von Nomen (Noun Non-animacy*; NN*).

$$NA = \frac{\#NAS}{\#NAS + \#NNAS} \qquad NN = \frac{\#NNAS}{\#NAS + \#NNAS}$$

Abb. 47 Formeln für die Berechnung von Belebtheitsinformationenvon [Evans et al. 2000:156]

Für NPs in der Subjektposition werden ähnliche Kennzahlen anhand der Verben berechnet. [Evans et al. 2000:159] nennen diese „Verb Subject Animacy (*VSA*)“ bzw. „Verb Subject Non-Animacy *(VSN)*.“

Im Weiteren wird überprüft, ob die NP ein Reflexivpronomen oder das Relativpronomen *who* enthält oder ob die NP ein genusmarkiertes Pronomen ist. Anhand einiger Heuristiken und des Algorithmus, der in Abbildung 47 aufgezeigt wird, kann nun die Belebtheit bestimmt werden. So werden Monatsnamen und Abkürzungen von Monatsnamen als unbelebt klassifiziert. Alle übrigen grossgeschriebenen Nomen werden als belebt und die übrigen NPs als unbelebt eingestuft. t_1, t_2 und t_3 sind Schwellenwerte.

```
NP is the noun phrase for which animacy has to be determined
Compute NA, NN, VSA, VSN for NP

  1.  If NA > t1 then NP is animate; Stop; Endif
  2.  If NN >  t2 then NP is inanimate; Stop, Endif
  3.  If NA>NN and VSA>VSN then NP is animate; Stop; Endif
  4.  If (NP contains the complementisers who or a reflexive) or (VSA > t3) then NP is animate; Stop;
      Endif
  5.  If unique_beginner (NP) = 15 or
        unique_beginner (NP) = 17 then
        NP is inanimate; Stop;
        // the unique beginners 15 and 17 from WordNet refer
        // to object and locations
      Endif
  6.  If head (NP) does not contain vowels then
        NP is inanimate; Stop;
        // this rule eliminates acronyms
      Endif
  7.  If head (NP) is in a first name gazetteer then
        NP is animate; Stop;
      Endif
  8.  NP is inanimate
```

Abb. 48 Algorithmus von [Evans et al. 2000:160] zur Ermittlung von Belebtheitsinformationen

[Evans et al. 2000] baut den Algorithmus zur Bestimmung von Belebtheitsinformationen in das AR-System MARS ein. Der Algorithmus erreicht laut [Evans et al. 2000:160] einen Präzisionswert von 77,25% und eine Ausbeute von 75.88%. Mit dem eingebauten Algorithmus kann die Anzahl der aufgelösten Pronomen von 203 auf 224 gesteigert werden.
Es muss allerdings beachtet werden, dass in einem Text nicht nur Personennamen, sondern auch Firmennamen und Produktenamen grossgeschrieben werden, welche als unbelebte Einheiten gekennzeichnet werden müssen. Auch muss der Algorithmus auf unterschiedlichen Textgenres getestet und die Schwellenwerte überprüft werden. Ausserdem bedarf es eines Namenerkennungssystems zur Verbesserung des Algorithmus. Für technische Texte ist dieser Ansatz aber weniger geeignet, da Fachwörter aus der Technik oft nicht in WordNet gefunden werden und bestimmte Begriffe spezielle Bedeutungen haben. Wie [Mitkov et al. 2002b:8] schreibt, wird beispielsweise das Nomen *printer* in WordNet mit belebten Einheiten in Verbindung gebracht. In technischen Texten handelt es sich aber um unbelebte Einheiten.

Nachdem die Schwierigkeiten in der AR und aktuelle Lösungsansätze zur automatischen Behandlung von Anaphern aufgezeigt worden sind, werden im Kapitel 6 die wichtigsten Erkenntnisse nochmals zusammengefasst.

6 Fazit

Die AR hat eine Schlüsselfunktion in zentralen Gebieten der automatischen Sprachverarbeitung wie der Maschinellen Übersetzung, der Textzusammenfassung und der Informationsextraktion. Seit den 60er Jahren ist AR ein Forschungsgebiet und insbesondere in den letzten 20 Jahren wurden viel versprechende Arbeiten vorgestellt. Während die früheren Arbeiten extensiv Sprachwissen, Domänewissen und Weltwissen auszubeuten versuchten, konzentrieren sich die zeitgenössischen Ansätze auf wissensarme Strategien, die auch für grosse Textmengen kostengünstig implementierbar sind.

Fortschritte in der AR sind auch auf das Aufkommen von Ontologien, Korpora und des Webs zurückzuführen, aus denen die zur Auflösung benötigte Information gewonnen werden kann und die auch zu Trainings- und Evaluationszwecken verwendet werden können. Auch Techniken des MLs und bilinguale Systeme zur Behandlung von Anaphern scheinen Verbesserungen mit sich zu bringen. Trotz den gewaltigen Fortschritten im Bereich der AR gibt es aber zahlreiche Schwierigkeiten, die noch zu bewältigen sind.

Einerseits ist die Performanzverbesserung von AR-Systemen ein zentraler Punkt. Dies bedingt einerseits Verbesserungen in der Vorverarbeitung. Neben dem Parsen und der Namenserkennung (Produktnamenerkennung inbegriffen) müssen auch die Identifikation von pleonastischen Pronomen und nicht-anaphorischen NPs und die Ermittlung von Belebtheits- und Genusinformation verbessert werden. Zudem sind auch Verbesserungen der Ressourcen notwendig. Die Informationsgewinnung aus Ontologien und Thesauri ist noch nicht optimal und es stehen den Forschern nur wenige umfassend annotierte Korpora zur Verfügung. Wichtig ist insbesondere die Annotation von anaphorischen Beziehungen, deren Einheitlichkeit noch nicht gewährleistet ist. Aufgrund der unzureichenden Ergebnisse in der Vorverarbeitungsphase operieren nur sehr wenige AR-Systeme in einem vollautomatischen Modus und manuelle Eingriffe sind an der Tagesordnung.

Andererseits stellt sich die Frage, wie weit die Performanz von AR-Systemen gesteigert werden kann und wo man mit wissensarmen Ansätzen an die Grenzen kommt, zumal das Fehlen von semantischem Wissen, Domänewissen und Weltwissen mit enormen Einschränkungen einhergeht. Semantisches Wissen in Form von Auswahlrestriktionen, die aus Ontologien und Thesauri gewonnen werden können, scheint unumgänglich in der Performanzsteigerung von wissensarmen Ansätzen. Weltwissen und Kontextwissen für die Disambiguierung von Anaphern ist aber kaum implementierbar, insbesondere dann wenn nicht nur bestimmte Textgenres berücksichtigt werden. Weitere Analysen der Auswahl- bzw. der Einschränkungskrite

rien könnten aber zu Verbesserungen im Auflösungsprozess führen, da bestimmte Faktoren nicht nur genreabhängig sind, sondern auch voneinander abhängen.

Der heutige Wissensstand erlaubt eine robuste, halbautomatische Behandlung von Pronomen. Komplexere Anaphern wie vollständige NPs wurden zwar in Angriff genommen, ihre automatische Behandlung beschränkt sich aber auf bestimmte anaphorische Beziehungen, wie Synonymie-, Hyperonymie oder Teil-Ganzes-Beziehungen. Anaphern noch komplexerer Art wie Adverb und VP-Anaphern scheinen ganz ausser Acht gelassen worden zu sein. Dies ist allerdings nicht verwunderlich, da man, wie in dieser Arbeit verdeutlicht wurde, bereits bei der Behandlung von Proformen und vollständigen NPs auf genügend Schwierigkeiten stösst.

Die Schwierigkeiten in der Erkennung und Ausfilterung von pleonastischen Pronomen und nicht-anaphorischen Einheiten sowie die Bestimmung eines Antezedens einer Anapher sind noch nicht befriedigend gelöst und auch für die Problematik der Diskrepanzen in den Kongruenzmerkmalen gibt es nur unzureichende Lösungsansätze.

In den heutigen AR-Ansätzen hat die Verwendung von Heuristiken einen hohen Stellenwert. Es ist aber schwierig Heuristiken aufzustellen, die nicht genreabhängig. Zudem ist es auch schwierig, Suchmuster zu generieren, die weder zu einschränkend noch zu allgemeingültig sind, und auch die Vollständigkeit von Listen ist eine kaum lösbare Aufgabe.

Der maschinellen Verarbeitung von Anaphora eröffnen sich daher noch etliche Fragen, insbesondere in der Behandlung der verschiedenen Typen von Anaphern und der einheitlichen Evaluation der Systeme. Die Anaphernresolution wird deshalb ein beliebtes Forschungsgebiet in der Computerlinguistik bleiben.

7 Literaturverzeichnis

[Ariel 1988] Ariel Mira, "Referring and accessibility," *Journal of Linguistics*, Vol. 24, No. 1, Cambridge University Press, pp. 65-87, 1983.

[Ariel 1996] Ariel Mira, "Referring expressions and the +/- coreference distinction," in: *Thorstein Fretheim and Jeanette K. Gundel (eds.): Reference and Referent Accessibility,* Benjamins, pp. 13-35, Amsterdam, Netherlands, 1996.

[Azzam et al. 1998] Azzam Saliha, Kevin Humphreys, Robert Gaizauskas, "Extending a Simple Coreference Algorithm with a Focusing Mechanism," in: *Botley Simon and Tony Mc Enery (eds.): New Approaches to Discourse Anaphora: Proceedings of the Second Colloquium on Discourse Anaphora and Anaphor Resolution (DAARC2)*, pp.15-27, Lancaster, UK, 1998.

[Stand: 22.8.2005]
www.dcs.shef.ac.uk/research/groups/nlp/external_papers/papers/daarc98.ps

[Barbu et al. 2001] Barbu Catalina and Ruslan Mitkov, "Evaluation tool for rule-based anaphora resolution methods," in: *Proceedings of ACL'01*, Toulouse, France, 2001.

[Stand:22.8.2005] http://acl.ldc.upenn.edu/P/P01/P01-1006.pdf

[Barbu et al. 2002] Barbu Catalina, Richard Evans, and Ruslan Mitkov, "A corpus based investigation of morphological disagreement in anaphoric relations," in: *Proceedings of LREC-2002 – Third International Conference on Language Resources and Evaluation,* Las Palmas de Gran Canaria, Spain, 2002.

[Stand:22.8.2005] http://clg.wlv.ac.uk/papers/barbu-2002-lrec-morph.pdf

[Balzert 1999] Balzert Helmut, *Lehrbuch Grundlagen der Informatik: Konzepte und Notationen in UML, Java, C++, Algorithmik und Software-Technik Anwendungen*, Spektrum, Heidelberg, Germany, 1999.

[Barlow 1998] Barlow Michael, "Feature Mismatches and Anaphora Resolution", in: *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'2),* pp. 34-41, Lancester, UK, 1998.

[Bean et al. 1999] Bean David L. and Ellen Riloff, "Corpus-Based Identification of Non-Anaphoric Noun Phrases," in: *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL'99)*, pp. 373-380, University of Maryland, USA, 1999.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/P/P99/P99-1048.pdf

[Bosch 2001] Bosch Peter, "Against the Identification of Anaphora and Presupposition," in: *J. van Kuppevelt and R. Smith (eds.): Proceedings of the 2nd SIGdial Workshop on Discourse and Dialogue*, Aalborg, Denmark, 2001.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/W/W01/W01-1604.pdf

[Bunescu 2003] Bunescu Razvan, "Associative Anaphora Resolution: A Web-Based Approach," in: *R.Dale, K. van Deemter, and R.Mitkov, (eds.): Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, pp. 47-52, Budapest, Hungary, 2003.

[Stand: 22.8.2005] http://www.itri.bton.ac.uk/~Kees.van.Deemter/bunescu.pdf

[Byron 1999] Byron Donna K., "Resolving Pronominal Reference to Abstract Entities," Thesis Project Proposal, URCS Technical Report No. 714, University of Rochester, Rochester, UK, 1999.

[Stand:22.8.2005] https://urresearch.rochester.edu/handle/1802/728

[Byron 2001] Byron Donna K., "The Uncommon Denominator: A Proposal for Consistent Reporting of Pronoun Resolution Results," in: *Computational Linguistics*, Vol. 27, No. 4, MIT Press, pp. 569-578, Cambridge, UK, 2001.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/J/J01/J01-4006.pdf

[Byron 2002] Byron Donna K., "Resolving Pronominal Reference to Abstract Entities," in: *Proceedings of the 2002 annual meeting of the Association for Computational Linguistics (ACL2002)*, pp. 80-87, Philadelphia, USA, 2002.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/P/P02/P02-1011.pdf

[Byron 2004a] Byron Donna K., *Resolving Pronominal Reference to Abstract Entities*, Technical Report No. 815, University of Rochester, Rochester, UK, 2004.

[Stand: 22.8.2005] http://www.cse.ohio-state.edu/~dbyron/papers/thesis.pdf

[Byron et al. 2004b] Byron Donna K. and Whitney Gegg-Harrison, "Eliminating Non-Referring Noun Phrases from Coreference Resolution," in: *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2004)*, pp. 21-26, S. Miguel, Azores, Portugal, 2004.

[Stand:22.8.2005] http://www.cse.ohio-state.edu/~dbyron/DAARC2004-Byron-Geggharr-app.pdf

[BNC] British National Corpus.

[Stand:22.8.2005] http://es-otto.unizh.ch/cgi-binbnc2/BNCquery.pl?urlTest=yes

[Cardie et al. 2002] Cardie Claire and Vincent Ng, "Identifying Anaphoric and Non-Anaphoric Noun Phrases to improve Coreference Resolution," in: *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, 2002.

[Stand:22.8.2005] http://www.cs.cornell.edu/home/cardie/papers/coling02.pdf

[Clark 1977] Clark, Herbert H., "Bridging," in: *P.N. Johnson-Laird and P.C. Wason (eds.): Thinking: Readings in Cognitive Science,* Cambridge University Press, pp. 169-174, London, 1977.

[Stand: 22.8.2005] http://ucrel.lancs.ac.uk/acl/T/T75/T75-2034.pdf

[Cristea 2001] Cristea Dan, "Anaphora", *EUROLAN-2001*, Iasi, Romania, 2001.

[Stand:22.8.2005] http://www.racai.ro/EUROLAN2001/page/resources/profs/Cristea/Anaphora.ppt#16

[Cristea et al. 2002] Cristea Dan, Oana-Diana Postolache, Gabriela-Eugenia Dima and Catalina Barbu, "AR-Engine – a framework for unrestricted co-reference resolution," in: *Proceedings of the LREC 2002 – Third International Conference on Language Resources and Evaluation*, Vol. 4, pp. 2000-2007, Las Palmas de Gran Canaria, Spain, 2002.

[Stand:22.8.2005] http://www.coli.uni-saarland.de/~oana/publications/lrec02.pdf

[Cristea et al. 2005] Cristea Dan and Oana-Diana Postolache, "How to Deal with Wicked Anaphora?," in: *Branco, António, Tony McEnery and Ruslan Mitkov (eds.): Anaphora Processing,* pp. 17-46, Amsterdam, Netherlands, 2005.

[Dagan et al. 1990] Dagan Ido and Alon Itai, "Automatic Processing of Large Corpora for the Resolution of Anaphora References," in: *Proceedings of the 13th International Conference on Computational Linguistics (COLING'90)*, Vol. 3, pp. 1-3, Helsinki, Finnland, 1990.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/C/C90/C90-3063.pdf

[Dale et al. 2005] Dale Robert, Hwee Tou Ng, Yu Zhou and Mary Gardiner, "Machine Learning Approach to Identification and Resolution of *One*-Anaphora," in: *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, Edinburgh, Scotland, 2005.

[Stand: 22.8.2005] http://www.comp.nus.edu.sg/~nght/pubs/ijcai05_1anaphor.pdf

[Denber 1998] Denber Michel, *Automatic Resolution of Anaphora in English,* Internal Report, Eastman Kodak Co., 1998.

[Dorrepaal 1990] Dorrepaal Joke, *Discourse Anaphora*, Utrecht, Netherlands, 1990.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/C/C90/C90-2017.pdf

[Eschenbach et al. 1989] Eschenbach Carola, Christopher Habel, Michael Herweg, Klaus Rehkämper, "Remarks on Plural Anaphora," in: *Proceedings of the Forth Conference of the European Chapter of the Association for Computational Linguistics,* Hamburg, 1989.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/E/E89/E89-1022.pdf

[Evans et al. 2000] Evans Richard and Orasan Constantin, "Improving anaphora resolution by identifying animate entities in texts," in: *Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC 2000)*, pp. 154-162, Lancaster, UK, 2000.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/evans-DAARC-2000.pdf

[Evans 2001] Evans Richard, "Applying Machine Learning Toward an Automatic Classification of *It*," *Literary and Linguistic Computing,* Vol. 16, No. 1, pp. 45-57, 2001.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/evans-01a.pdf

[Evans 2002] Evans Richard, "Refined Salience Weighting and Error Analysis in Anaphora Resolution," in: *Proceedings of RRNLP-2002*, pp. 51-59, University of Alicante, Alicante, Spain, 2002.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/evans-02.pdf

[Ferrández et al. 1998] Ferrández Antonio, Manuel Palomar, Lidia Moreno, "A computational approach to pronominal anaphora, one-anaphora and surface-count anaphora," in: *Proceeding of the Discourse Anaphora and Resolution Colloquium*, pp. 117-128, 1998.

[Gasperin et al. 2004] Caroline Gasperin, Susanne Salmon-Alt, Renata Vieira, "How useful are similarity word lists for indirect anaphora resolution?," in: *Proceedings of DAARC 2004*, Sao Miguel, Azores, 2004.

[Stand: 22.8.2005] http://www.cl.cam.ac.uk/users/cvg20/daarc04.pdf

[Gundel et al. 2005] Gundel, Jeanette K., Nancy A. Hedberg and Ron Zacharski, "Pronouns Without NP Antecedents: How do we Know when a Pronoun is Referential?," in: *Branco, António, Tony McEnery and Ruslan Mitkov (eds.): Anaphora Processing,* pp. 351-364, Amsterdam, Netherlands, 2005.

[Stand: 22.8.2005] http://www.zacharski.org/papers/GHZ-Lisbon.pdf

[Halliday et al. 1976] Halliday, M. A. K. and Ruqaiya Hassan, *Cohesion in English,* Longman, London, UK, 1976.

[Jurafsky et al. 2000] Jurafsky, Daniel, and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, Prentice-Hall, New Jersey, USA, 2000.

[Kehler et al. 2004] Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma, "Competitive Self-Trained Pronoun Interpretation," in: *Proceedings of HLT/NAACL 2004*, Boston, MA, USA, 2004.

[Stand: 22.8.2005] http://ling.ucsd.edu/~kehler/Papers/KehlerEtAl-NAACL-2004-short.pdf

[Lappin et al. 1990] Lappin Shalom and Michael McCord, "Anaphora Resolution in Slot Grammar," in: *Computational Linguistics*, Vol. 16, No. 4, MIT Press, pp. 197-212, 1990.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/J/J90/J90-4001.pdf

[Lappin et al. 1994] Lappin Shalom and Herbert J. Leass, "An Algorithm for Pronominal Anaphora Resolution," *Computational Linguistics*, Vol. 20 No. 4, pp. 535-561, 1994.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/J/J94/J94-4002.pdf

[Lappin et al. 1996] Lappin Shalom and Hsue-Hue Shih, "A generalized reconstruction algorithm for ellipsis resolution," in: *Proceedings of COLING-96*, pp. 351-364, University of Copenhagen, Copenhagen, Denmark.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/C/C96/C96-2116.pdf

[Linke et al. 2004] Linke Angelika, Markus Nussbaumer und Paul R. Portmann, *Studienbuch Linguistik. 5., erweiterte Auflage,* Niemeyer, Tübingen, Germany, 2004.

[Litrán et al. 2004] Litrán José Carlos Clemente, Kenji Satou and Kentaro Torisawa, "Improving the Identification of Non-Anaphoric *it* using Support Vector Machines," *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP 2004)*, Geneva, Switzerland, 2004.

[Stand: 22.8.2005] http://www.tori.jaist.ac.jp:8000/papers/clemente_coling04.pdf

[Markert et al. 2003] Markert Katja, Malvina Nissim, and Natalia N. Modjeska, "Using the Web for nominal anaphora resolution," in: *Robert Dale, Kees van Deemter, and Ruslan Mitkov (eds.): Proceedings of the EACL Workshop on the Computational Treatment of Anaphora*, 2003.

[Stand: 22.8.2005] http://www.itri.bton.ac.uk/~Kees.van.Deemter/markert.pdf

[Meyer et al. 2002] Meyer Josef and Robert Dale, "Using the WordNet Hierarchy for Associative Anaphora Resolution," in: *Proceedings of SemaNet'02: Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/W/W02/W02-1112.pdf

[Mitkov 1997] Mitkov Ruslan, "Factors in anaphora resolution: they are not the only things that matter. A case study based on two different approaches," in: *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution*, pp. 14-21, Madrid, Spain, 1997.

[Stand: 22.8.2005]
http://people.csail.mit.edu/u/j/jrennie/public_html/papers/other/mitkov-coreference-97.pdf

[Mitkov 2001] Mitkov Ruslan, "Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems," *Applied Artificial Intelligence: An International Journal*, Vol. 15, No. 3, pp. 253-276, 2001.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/Lancaster2000.PDF

[Mitkov 2002a] Mitkov Ruslan, *Anaphora Resolution*, Longman, London, UK, 2002.

[Mitkov et al. 2002b] Mitkov Ruslan, Richard Evans and Constantin Orasan, "A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method," in: *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, 2002.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/ciclingAR19.pdf

[Mitkov et al. 2002c] Mitkov Ruslan and Catalina Barbu, "Using bilingual corpora to improve pronoun resolution," *Languages in Contrast,* Vol 4, No. 2*,* Benjamins, pp. 201-211, 2002.

[Stand: 22.8.2005] http://clg.wlv.ac.uk/papers/mitkov02.pdf

[Modjeska 2002] Modjeska Natalia N., "Lexical and Grammatical Role Constraints in Resolving *Other*-Anaphora," in: *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, pp. 129-134, Lisbon, Portugal, 2002.

[Stand: 22.8.2005] http://ftp.cs.toronto.edu/pub/gh/modjeska-daarc2002.pdf

[Modjeska 2003] Modjeska Natalia N., "Resolving Other-Anaphora", Ph.D. Thesis, University of Edinburgh, Edinburgh, Scottland, 2003.

[Stand: 22.8.2005] http://www.cs.toronto.edu/pub/gh/modjeska-f-thesis.pdf

[Mollá et al. 2003] Mollá Diego, Rolf Schwitter, Fabio Rinaldi, James Dowdall und Michael Hess, "Anaphora Resolution in ExtrAns," in: *ARQAS, Anaphora Resolution in Question Answering and Summarization, Proceedings of 2003 International Symposium on Reference Resolution and its Applications to Question Answering and Summarization*, pp. 67-74, Venice, Italy, 2003.

[Stand: 22.8.2005] http://www.ics.mq.edu.au/~rolfs/papers/extrans_anaphora.pdf

[Paice et al. 1987] Paice, Chris D. and G.D. Husk, "Towards an automatic recognition of anaphoric features in English text: the impersonal pronoun 'it'," *Computer speech and Language*, Vol. 2, Academic Press, pp. 109-132, USA, 1987.

[Peral et al. 2000] Peral Jesús and Antonio Ferrández, "An Application of the Interlingua System ISS for Spanish-English Pronominal Anaphora Generation," *NAACL-ANLP 2000 Workshop: Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, pp. 42-51, Seattle, Washington, USA, 2000.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/W/W00/W00-0206.pdf

[Poesio et al. 2002] Poesio Massimo, Tomonori Ishikawa, Sabine Schulte im Walde, and Renata Vieira, "*Acquiring lexical knowledge for anaphora resolution*," in: *Proceedings of LREC 2002, – Third International Conference on Language Resources and Evaluation,* pp. 1220-1224, Las Palmas de Gran Canaria, Spain, 2002.

[Stand: 22.8.2005] http://cswww.essex.ac.uk/poesio/publications/LREC02.pdf

[Poesio 2003] Poesio Massimo, "Associative Descriptions and Salience: A Preliminary Investigation," in: *Proceedings of the EACL Workshop on Computational Treatments of Anaphora*, Budapest, Hungary, 2003.

[Stand: 22.8.2005] http://www.itri.bton.ac.uk/~Kees.van.Deemter/poesio.pdf

[Poesio et al. 2004] Massimo Poesio, Rahul Mehta, Axel Maroudas and Janet Hitzeman, "Learning to resolve bridging references," in: *Proceedings of ACL*, Barcelona, Spain, 2004.

[Stand: 22.8.2005] http://cswww.essex.ac.uk/staff/poesio/publications/ACL04.pdf

[Postolache 2004] Postolache Oana. "RARE-Robust Anaphora Resolution Engine," Master Thesis, University of Iasi, Iasi, Romania, 2004.

[Stand: 22.8.2005] http://www.coli.uni-sb.de/~oana/publications/Postolache-MSc-Thesis.pdf

[Prince 1992] Prince, Ellen F., "The ZPG Letter: Subjects, Definiteness, and Informationstatus," in: *Thompson, S. and W. Mann (eds.)*: *Discourse Description: diverse analyses of a fund raising text*, Benjamins, pp. 295-325, Amsterdam, Netherlands, 1992.

[Stand: 22.8.2005] http://www.coli.uni-saarland.de/~korbay/page.php?id=helsinki05

[Sayed 2003] Imran Q. Sayed, "Issues in Anaphora Resolution", 2003.

[Stand: 22.8.2005] http://nlp.stanford.edu/courses/cs224n/2003/fp/iqsayed/project report.pdf

[Schilder 1999] Schilder Frank, „Pointing to Events," in: *Proceedings of the Ninth Conference of the European Chapter of the Association of Computational Linguistics (EACL)*, pp. 283-285, University of Bergen, Norway, 1999.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu/E/E99/E99-1049.pdf

[Siddharthan 2003] Siddharthan Advaith, "Resolving Pronouns Robustly: Plumbing the Depths of Shallowness," in: *Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pp. 7-14, Budapest, Hungary, 2003.

[Stand: 22.8.2005] http://www1.cs.columbia.edu/~as372/anaphora_EACL03.pdf

[Strüber 2001] Strüber Frank, "Eine Link-Grammatik für einen Ausschnitt des Deutschen: Studienarbeit im Fach Informatik," Friedrich-Alexander-Universität, Erlangen-Nürnberg, Germany, 2001.

[Stand: 22.8.2005] http://www2.informatik.uni-erlangen.de/Lehre/SA-DA/old/SA-abgeschlossen/strueber.pdf

[Tanaka 1999] Tanaka Izumi, "The value of an annotated corpus in the investigation of anaphoric pronouns, with particular reference to backwards anaphora in English," Ph.D. Thesis, Lancaster University, Lancaster, 1999.

[Vieira et al. 2000] Vieira Renata and Massimo Poesio, "An Empirically Based System for Processing Definite Descriptions," *Computational Linguistics*, Vol. 26, No. 4, MIT Press, pp. 539-594, 2000.

[Stand: 22.8.2005] http://acl.ldc.upenn.edu//J/J00/J00-4003.pdf

[Wilson et al. 1998] Wilson Deirdre and Tomoko Matsui, "Recent Approaches to Bridging: Truth, coherence, relevance," *UCL Working Papers in Linguistics 10,* pp. 173-200, 1998.

[Stand: 22.8.2005] http://www.phon.ucl.ac.uk/home/PUB/WPL/98papers/wilson.pdf

[Wind 2002] Wind Lutz, „Maschinelle Analyse des Textzusammenhanges (Bridging Phänomene)," Magisterarbeit, Ruprecht Karls Universität Heidelberg, Heidelberg, Germany, 2002.

[Stand: 22.8.2005] http://www.computational-linguistics.de/docs/MA_bridging5.pdf

8 Glossar

Begriff	Bedeutung/Beispiel
Adjunkt	grammatikalisch nicht notwendige Ergänzung (z.B. *She watched TV **every day***)
Adverbanahpher	als Adverb realisierte Anapher (z.B. *there*)
Algorithmus	Vorgang mit Handlungsanweisungen zur Lösung eines Problems
alinierter Korpus	Korpus, in dem die expliziten Entsprechungen zwischen der Ausgangssprache und der Übersetzung in die Zielsprache markiert werden (z.B. durch Darstellung in Spalten)
Anapher	rückverweisender Ausdruck
Anaphernresolution	Bestimmung des Antezedens einer Anapher
anaphorisch	auf bereits erwähnte Einheit verweisend
Antezedens	Bezugswort im Text
Apposition	nachträgliche Ergänzung (durch Kommata abgegrenzt)
Argument	Ein Prädikat erfordert zu seiner Spezifikation bestimmte Argumente (*Peter, Sarah, an apple* in *Peter gave Sarah an appl**e***)
assoziative Anapher	Der Referent einer assoziativen Anapher ist nicht explizit im Text erwähnt. Die Anapher kann aber über das Antezedens verstanden werden. Anapher und Antezedens sind aber nicht referenzidentisch.
Attribut	ein Attribut ordnet einer Einheit ein Merkmal zu
attributiv	beifügend
Ausbeute (Recall)	Evaluationsmesswert; Quotient der Anzahl der korrekt aufgelösten Anaphern und der Anzahl der aufzulösenden Anaphern
Computerlinguistik, computerlinguistisch	Disziplin, die sich mit der maschinellen Verarbeitung der natürlichen Sprache beschäftigt
Deixis, deiktisch	bei deiktischen Verweisen bezieht sich ein Ausdruck nicht auf den Diskurs, sondern auf die Sprechersituation (z.B. *ich, du , hier*); Verweise mit Gesten auf Gegenstände, Orte, Ereignisse etc. der realen Welt
direkte Anapher	Anapher, die direkt aufgelöst werden kann; Referent ist im Text erwähnt
Diskurs	hier: Text und Dialog gemeint
Diskurseinheit	Einheit aus einem Text oder einem Dialog
Endophora, endophorisch	Verweis auf textinterne Einheiten

Ereignis-Anapher	Anapher, die sich auf ein Ereignis, eine Situation, eine Aktivität, ein Faktum, eine Tatsache, eine Proposition oder eine derartige Einheit bezieht
Erfolgsrate (nicht-triviale/kritische)	Quotient der Anzahl der korrekt aufgelösten Anaphern und der Anzahl aller Anaphern. Die nicht-triviale Erfolgsrate beinhaltet Anaphern, die mehr als einen Antezedenskandidaten haben. Die kritische Erfolgsrate enthält Anaphern, die nach der Anwendung von Genus- und Numerusfiltern mehr als einen Antezedenskandidaten haben.
Evaluation	Auswertung
Exophora, exophorisch	Verweis auf textexterne Elemente
generisch	das Geschlecht oder die Gattung betreffend
genetischer Algorithmus	Optimierungsverfahren basierend auf Heuristiken
Genus, Genera	Geschlecht (männlich, weiblich, sächlich)
Gerundium	Form, die verwendet wird, um aus einem Verb ein Substantiv zu machen. Im Englischen ist es die *-ing*-Form.
Heuristik	Verfahren zur Problemlösung, das auf Hypothesen, Analogien und/oder Erfahrungen aufgebaut ist.
Hyperonymie, Hyperonym,	Überordnung; Hyperonym ist ein Oberbegriff (z.B. „Hund“ ist Hyperonym von „Dogge“);
Hypernymie, Hypernym	Unterordnung; Hypernym ist ein Unterbegriff (z.B. „Dogge“ ist Hypernym von „Hund“)
Idiom, idiomatisch	feste Wortverbindung; aus der Bedeutung der Wörter kann nicht auf die Gesamtbedeutung geschlossen werden (z.B. *to kick the bucket*)
Intersententielle Anapher	satzübergreifende Anapher; Anapher und Antezedens nicht im selben Satz
intrasententiel, intrasententielle Anapher	Anapher und Antezedens im selben Satz
Katapher, kataphorisch	vorausverweisender Ausdruck
kognitive Verben	Verben der Wahrnehmung
kollektive Nomen	Nomen, die Gruppierungen von Leuten, Tieren, Objekten etc. beinhalten (z.B. *a flock of sheep*)
Kollokation	häufig auftretende Wortverbindung
Kompositum	Einheit, die aus eigenständigen Teilwörtern zusammengesetzt ist
Kongruenzmerkmale	grammatikalische Übereinstimmungsmerkmale
Kopula	Konstrukte, in denen ein Prädikat dem Subjekt zugesprochen wird oder das Subjekt ist (z.B. *He is a teacher*)
Koreferenz, koreferent	verweisender Ausdruck und sein Bezugswort beziehen sich auf die selbe aussersprachliche Einheit

Korpus, Korpora	umfassende Textsammlung
Lemma, Lemmata	Grundform eines Wortes, die im Lexikon steht
LINK-Grammatik	Die Grundidee der LINK-Grammatik besteht darin, jedem Wort mögliche Konnektoren zuzuordnen, die seine Bindungsmöglichkeiten beschreiben. Jede Verbindung (LINK) von zwei Wörtern hat einen Namen, der die Beziehung zwischen den Wörtern angibt. So trägt beispielsweise der Link, der das Subjekt und das Verb verbindet, den Namen S.
Maschinenlernen (ML)	automatische Akquisition von Wissen aus Trainingsdaten
Meronymie, Meronym	Teil-Ganzes-Beziehung; „Tischbein“ ist ein Meronym von „Tisch“
Modales Adjektiv	Adjektiv, das nicht prädikativ, nicht graduierbar und nicht morphologisch abgeleitet ist (z.B. *a* ***total*** *idiot*)
Modalverb	Hilfsverb
Modifikator	sprachliche Einheit, die eine andere sprachliche Einheit näher bestimmt (z. B. *a* ***blue*** *skirt*)
Mustervergleich (Pattern Matching)	Zeichenketten werden auf ihre Ähnlichkeit zueinander geprüft
nicht-anaphorisch	nicht auf bereits erwähnte Einheit verweisend
Nomen	Substantiv, Hauptwort
Nominalphrase (NP)	Satzbestandteil; die Zusammensetzung eines Nomens mit den dazugehörenden Wörtern wie Artikeln, Adjektiven etc. zu einer Phrase
NP-Anapher	Anapher, die als NP realisiert ist. Dazu gehören Pronomen, *One*-Anaphern, Eigennamen, definite und indefinite NPs, und *other*-Anaphern.
Nullanapher	Anapher, die sich als Lücke manifestiert
Numerus	Zählform in der Grammatik (Ein- und Mehrzahl)
***One*-Anapher**	als *one* realisierte Anapher
Ontologie	Wissensrepräsentation; eine hierarchische Ordnung von Begriffen
Parser	Programm zur Ermittlung von Syntaxstrukturen eines Satzes
Pattern Matching	Mustervergleich
Performanz (F-Measure)	Evaluationsmesswert, der sich aus der Präzision und Ausbeute zusammensetzt
pleonastisches Pronomen	*it*-Pronomen, das kein Antezedens hat (z.B. ***it*** *is raining*)
POS (Part of Speech), POS-Tag	Wortklasse; eine POS-Tag, ist eine Markierung, die die Wortklasse beinhaltet
POS-Tagger	Tagger, der jedem Wort eine Wortklasse zuordnet
Postmodifikator	Modifikator, der nach der zu bestimmenden Einheit steht
Prädikat	Element, das andere Elemente (sogenannte Argumente) zur Spezifikation benötigt (z.B. *Peter* ***gave*** *Sarah* *an apple*)

Pragmatik, pragmatisch	Teildisziplin der Sprachwissenschaft, die sich mit der Verwendung der Sprache befasst. Man kann Sprache verwenden, um beispielsweise jemanden zu loben oder jemandem zu drohen.
Prämodifikator	Modifikator vor der zu bestimmenden sprachlichen Einheit
Präpositionalphrase (PP)	Phrase bestehend aus einer Präposition und einer NP; PPs sind auch Bestandteile von VPs und NPs
Präzision (Precision)	Evaluationsmesswert; Quotient der Anzahl der korrekt aufgelösten Anaphern und der Anzahl der aufgelösten Anaphern
probabilistische Ansätze	wahrscheinlichkeitsbasierte Ansätze
Proform	Verweisform im Text
Proposition	eigentlicher Satzinhalt eines Sprechaktes
Referent	aussersprachliches Objekt auf das Wörter im Diskurs verweisen; Einheit aus der realen Welt
referenzähnliche Anapher	Anapher und Antezedens haben einen bedeutungsähnlichen Referenten
referenzidentische Anapher	Anapher und Antezedens haben denselben Referenten in der realen Welt
reziproke Pronomen	wechselweise aufeinander bezügliche Pronomen (z.B. *each other*)
Semantik, semantisch	Teilgebiet der Linguistik, das sich mit der Bedeutung der sprachlichen Zeichen befasst
SLOT-Grammatik	In der SLOT Grammatik hat jedes Wort eine bestimmte Anzahl von freien Stellen. Regeln legen fest, wie diese Leerstellen gefüllt werden dürfen. So verlangen Verben und Substantive beispielsweise bestimmte Präpositionen (*wait for, opinion about*).
Synonymie, Synonym	Gleichheitsbeziehung; Wort mit sehr ähnlicher Bedeutung
Syntax, syntaktisch	Satzbau betreffend
Tagger	Programm, das Wörter mit Etiketten versieht (z.B. mit der Wortklasse)
Thesaurus	Wortnetz; Begriffe sind mit Relationen miteinander verbunden
Treebank	Korpus mit bereits geparsten Sätzen
Verbalphrase (VP)	Satzbestandteil. Phrase bestehend aus einem Verb und den dazugehörenden Nominalphrasen und/oder Präpositionalphrasen
VP-Anapher	Anapher, die als Verb realisiert ist (z.B. *did*)
WordNet	Wortnetz, in dem Zusammenhänge zwischen Wörtern beschrieben sind

Lebenslauf

Renata Sevcikova
Lochstr. 8
8200 Schaffhausen
Tel.: 078 860 27 97
e-mail: bijou@shlink.ch

27. 01.78
ledig
CH

Ausbildung

- Primarschule Breite, Schaffhausen 1984-1990
- Sekundarschule am Bach, Schaffhausen 1990-1994
- Kantonsschule Schaffhausen, Typus E 1994-1998
- Studium der Anglistik, Computerlinguistik, und Geographie, Universität Zürich 1998-2006

- Auslandaufenthalt in Neuseeland (9 Mte.), Proficiency 2001/2002
- Auslandaufenthalt in Kanada und den USA (3,5Mte) 2003

Berufliche Tätigkeit

- Jumbo Bau- und Freizeitmarkt AG, SH, Kassiererin und Allrounderin 1999-2000
- Kieser Training AG (SH, ZG, ZH), Instruktorin und Rezeptionistin 2000-2005
- Snowboardschule PAT, Stoos, Snowboardlehrerin 2002-2004
- Zürcher Kantonalbank, Zürich, Kundenberaterin 2004-2006

Hobbies

- Sport (Snowboarden, Windsurfen, Surfen, Laufsport, Fussball)
- Sprachen
- Reisen