Ludwig-Maximilians-Universität München
Centrum für Informations- und Sprachverarbeitung (CIS)

Magisterarbeit
im
Studiengang Computerlinguistik

Grammatik der Menschenbezeichner in biographischen Kontexten

Gekürzte Fasssung vom September 2007

Michaela Geierhos

März 2006

Betreuer der Arbeit:
Prof. Dr. Franz Guenthner

Inhaltsverzeichnis

Vorwort **7**

1 NER innerhalb biographischer Relationen in Nachrichten **9**
1.1 Begriffsklärung: Named Entity Recognition (NER) 9
1.2 Definition: Biographische Relationen 11
1.2.1 Persönliche Relationen 11
1.2.2 Öffentliche Relationen 12
1.2.3 Zufällige Relationen 12
1.3 Einschätzung der Thematik 13
1.3.1 Probleme und Chancen 13
1.3.2 Bewältigung der Aufgabe 14

2 Lokale Grammatiken **15**
2.1 Was sind lokale Grammatiken? 15
2.2 Warum werden lokale Grammatiken verwendet? 16
2.3 UNITEX – Ein System zur Anwendung lokaler Grammatiken 18
2.3.1 Textvorverarbeitung 18
2.3.2 DELA Wörterbücher 19
2.3.3 Prioritäten bei der Anwendung der Lexika 22
2.3.4 Mustererkennung und Konkordanzen 22

3 Zusammenfassung früherer Arbeiten **23**
3.1 Bootstrapping 23
3.1.1 Bootstrapping bei der Entwicklung lokaler Grammatiken [Gross, 1999] 23
3.1.2 Bootstrapping zur Erkennung von Nominalphrasen mit FSTs [Senellart, 1998b] 24
3.2 Lemmatisierung zusammengesetzter Zeiten im Englischen [Gross, 1998-1999] 25
3.3 Erkennung von Personenbezeichnungen 28
3.3.1 Erkennung von Eigennamen und Berufsbezeichnungen [Senellart, 1998a] 28
3.3.2 Erkennung von Personennamen in Zeitungstexten [Friburger, 2002] 34
3.4 Erkennung von Organisationsnamen in Wirtschaftsnachrichten [Mallchok, 2004] 37

4 Beschränkungen im System **41**
4.1 Sprachgebundenheit . . . 42
4.2 Schwerpunkt Wirtschaftsnachrichten . . . 42
4.3 Priorisierung von Entitäten . . . 43

5 Ressourcen: Grundlagen des Systems **45**
5.1 Korpora . . . 45
5.1.1 Financial Times . . . 45
5.1.2 Biography.com . . . 46
5.2 Wörterbuchressourcen . . . 47
5.2.1 Lexikon der Vornamen . . . 48
5.2.2 Lexikon der Nachnamen . . . 48
5.2.3 Lexika der Personennamen . . . 49
5.2.4 Lexika der Personentitel . . . 52
5.2.5 Lexika der allgemeinen Menschenbezeichner . . . 53
5.2.6 Lexikon der personenbezogenen Prädikate . . . 56
5.2.7 Lexika der Branchen . . . 57
5.2.8 Lexika der Organisationsnamen . . . 58
5.2.9 Lexika der geographischen Begriffe . . . 61
5.2.10 Lexika der Temporalia . . . 64
5.2.11 Weitere Lexika . . . 66
5.3 Verifikationsmöglichkeiten bei Google . . . 67

6 Grammatik der Menschenbezeichner **69**
6.1 Analyse von Personennamen . . . 69
6.1.1 Syntaktische Variabilität bei Personennamen . . . 69
6.1.2 Disambiguierung von „Scheinnamen“ . . . 72
6.1.3 Vervollständigung des Personennamenlexikons . . . 73
6.2 Allgemeine Menschenbezeichner . . . 74
6.3 Auflösen von Anaphern . . . 77

7 Grammatik der Organisationsnamen **79**
7.1 Syntaktische Variabilität bei Organisationsnamen . . . 79
7.2 Abgrenzung von unechten Organisationsnamen . . . 81
7.3 Vervollständigung des Organisationsnamenlexikons . . . 83

8 Grammatik der Ortsangaben **85**
8.1 Biographische Relationen mit Ortsangaben . . . 85
8.2 Ortsangaben in ihrer Funktion als Attribute . . . 86
8.2.1 Toponyme als Attribut einer Berufsbezeichnung . . . 86
8.2.2 Toponyme als Attribut eines Organisationsnamens . . . 86

9 Grammatik der Datumsangaben **89**

10 Grammatik persönlicher Relationen **97**
10.1 Die Geburt: „to be born“ . . . 98
10.2 Die Kindheit: „to be raised (up)“ . . . 100
10.3 Der Schulabschluss: „to graduate“ . . . 102
10.4 Die Heirat: „to be married“ . . . 105
10.5 Die Scheidung: „to be divorced“ . . . 108
10.6 Der Tod: „to die“ . . . 111

11 Grammatik beruflicher Relationen **115**
11.1 Der Beginn eines Beschäftigungsverhältnisses . . . 116
11.1.1 Die Ernennung: „to be appointed as“ . . . 116
11.1.2 Die Einstellung: „to employ so.“ . . . 122
11.1.3 Der Firmeneintritt: „to join“ . . . 124
11.2 Die Ausübung des Berufes . . . 124
11.2.1 Das Beschäftigungsverhältnis: „to be employed“ . . . 124
11.2.2 Die Bezahlung: „to be paid as“ . . . 125
11.2.3 Die Tätigkeit: „to work as“ . . . 126
11.3 Das Ende eines Arbeitsverhältnisses . . . 129
11.3.1 Die Entlassung: „to dismiss so.“ bzw. „to be dismissed“ . . . 129
11.3.2 Die Nachfolge: „to be replaced as“ . . . 131
11.3.3 Die Abdankung: „to resign as“ . . . 132
11.3.4 Die Pensionierung: „to retire so.“ bzw. „to be retired“ . . . 132

12 Auswertung der Ergebnisse **135**
12.1 Evaluationsmaße [Wikipedia, 2005/2006] . . . 135
12.1.1 Precision bzw. Genauigkeit . . . 135
12.1.2 Recall bzw. Vollständigkeit . . . 135
12.1.3 Fall-Out . . . 135
12.2 Qualität des Systems . . . 136

13 Anwendungen **137**
13.1 Extraktion von Relationen zwischen Personen und Organisationen . . . 137
13.2 Extraktion von Relationen zwischen mindestens zwei Personen . . . 138
13.3 Extraktion von Relationen zwischen Personen und ihren Berufen . . . 139

14 Zusammenfassung und Ausblick **141**

A Übersicht aller Kategorien in den Wörterbüchern **143**
A.1 Semantische Kategorien . . . 143
A.2 Grammatikalische Kategorien . . . 144

B Syntaktische Variabilität am Beispiel von „Bill Gates“ **145**

Literaturverzeichnis **147**

Index **154**

Vorwort

Wer muss heute noch eine Biographie oder einen Lebenslauf veröffentlichen, wenn er oder sie eine in der Öffentlichkeit präsente Person ist? – Eigentlich betrifft das niemand dieser Leute, denn Ausschnitte ihres Lebens werden in den Printmedien von verschiedensten Blickwinkeln beleuchtet. Natürlich veröffentlichen die wenigsten Zeitungen oder Magazine lückenlose Lebensläufe prominenter Menschen.

Liest man nur einen Artikel zu der betreffenden Person, so erfährt man nur wenig über sie und bekommt auch recht einseitige Informationen. Doch lässt man Google nach diesen bekannten Leuten suchen, so bekommt man eine Vielzahl von Artikelreferenzen, welche die unterschiedlichsten Facetten und Bereiche ihres öffentlichen und privaten Lebens beleuchten. Kurz und prägnant werden einem Informationen über den Familienstand, die Familienverhältnisse, den sozialen Status, das geschätzte oder bekannte Jahreseinkommen, sowie Vorlieben, Freizeitaktivitäten und noch vieles mehr auf dem „silbernen Tablett“ serviert. Die Fülle an Informationen, die Google ihren Kunden bietet, übersteigt oft ihre anfänglichen Erwartungen. Manchmal erfährt man sogar Details aus dem Leben der Reichen und Schönen, welche derjenige selbst wohl nie so veröffentlicht hätte.

Wie CNET News.com [Mills, 2005] am 14. Juli 2005 berichtete, erging es dem Google CEO Eric Schmidt nicht anders. Obwohl er selbst auf seiner Homepage wenig über seine Person preisgibt, findet Google nach kürzester Zeit alle wichtigen Daten, die seine Person betreffen.

> **Google CEO Eric Schmidt doesn't reveal much about himself on his home page.**
> *But spending 30 minutes on the Google search engine lets one discover that Schmidt, 50, was worth an estimated $1.5 billion last year. Earlier this year, he pulled in almost $90 million from sales of Google stock and made at least another $50 million selling shares in the past two months as the stock leaped to more than $300 a share.*
>
> *He and his wife Wendy live in the affluent town of Atherton, Calif., where, at a $10,000-a-plate political fund-raiser five years ago, presidential candidate Al Gore and his wife Tipper danced as Elton John belted out "Bennie and the Jets".*
>
> *Schmidt has also roamed the desert at the Burning Man art festival in Nevada, and is an avid amateur pilot.* [1]

[1] Ausschnitt aus dem Artikel *„Google balances privacy, reach“* von Elinor Mills [Mills, 2005]

Aber warum sollte man 30 Minuten bei der Suche nach „Google CEO Eric Schmidt“ damit verbringen, die einzelnen Treffer nach der biographisch relevanten Information zu durchsuchen? Wäre es nicht sinnvoller, wenn bei der Suche nach Personen auch ein Fokus auf die verschiedenen Beziehungen gelegt wird, in denen ein Mensch mit anderen Menschen, mit einer Firma, mit Wohn- und Arbeitsorten oder zeitlichen Begebenheiten in Verbindung steht? Würde es eine personenbezogene Suche nicht enorm erleichtern, wenn einer der allerersten Treffer Aufschluss über das Alter bzw. das Geburtsdatum, dann eventuell über den Familienstand, gefolgt vom Beruf oder dem momentanen Beschäftigungsverhältnis geben würde?

Damit will ich andeuten, dass eine Staffelung nach Wichtigkeit der biographischen Daten und das Ausfiltern von biographisch irrelevanter Information die Zufriedenheit des Benutzers bei der Suche deutlich erhöhen kann.

Doch bevor man eine Skala für die Relevanz von biographischen Relationen festlegen kann, muss man sich ein Bild davon machen, welche Prädikate überhaupt dafür in Frage kommen. Denn einerseits sollten es u.a. Verbrelationen sein, die sehr häufig in Biographien auftauchen, und andererseits müssen sie auch interessant für den Informationssuchenden sein.

Deshalb möchte ich im Rahmen dieser Magisterarbeit versuchen, eine umfassende Definition von Prädikaten zu geben, welche in biographischen Kontexten auftreten können und essentielle Informationen über die betreffende Person geben. Dabei verstehe ich unter einer „Definition von Prädikaten“ nicht nur eine reine Auflistung dieser, sondern vielmehr die Erstellung einer Grammatik – eines Regelwerkes – welche analysiert und gleichzeitig vorgibt, wie sich ein bestimmtes Verb innerhalb eines Satzgefüges verhält, d.h. welche Argumente das Verb zwingend oder optional hat, oder ob es oft im Zusammenhang mit Lokativa oder Temporalia auftritt. Natürlich ist diese syntaktische, aber auch semantische, Betrachtungsweise von personenbezogenen Sätzen sprachabhängig. Aufgrund der Vielzahl an möglichen Prädikaten, werde ich mich in meiner Arbeit ausschließlich auf eine einzige Sprache, nämlich das Englische, beschränken.

Dabei ist mir besonders wichtig, dass der Schwerpunkt dieser Arbeit nicht das Ranking von biographischen Relationen oder eine Fallstudie ist, wie eine gute, automatisch generierte Textzusammenfassung einer Biographie auszusehen hätte, sondern vielmehr möchte ich das Augenmerk auf die Analyse von biographisch relevanten Sätzen richten.

Es wird nicht bei einer reinen syntaktischen Studie von Satzgefügen bleiben, da die natürliche Sprache sehr viele Paraphrasierungsmöglichkeiten bietet. Das macht u.a. ein semantisches Clustering von Relationstypen – die Bildung von Synonymklassen auf der Ebene der Prädikate – aber auch eine Typisierung von Satzteilen notwendig. Letzteres macht bereits der Titel dieser Magisterarbeit deutlich, denn „Menschenbezeichner“ sind bereits eine eigene Klasse, worunter u.a. Eigennamen für Personen, wie z.B. *„Bill Clinton“*, Berufsbezeichnungen, wie z.B. *„software engineer“*, oder Bezeichnungen für Verwandtschaftsverhältnisse, wie z.B. *„mother, aunt, grandfather“*, fallen.

Aber bevor ich in Details gehe, sollte das als kleiner Vorgeschmack auf diese Arbeit ausreichend sein und natürlich hoffe ich, dass ich Ihr Interesse dafür geweckt habe.

Michaela Geierhos
München, den 27. März 2006

1 NER innerhalb biographischer Relationen in Wirtschaftsnachrichten

1.1 Begriffsklärung: Named Entity Recognition (NER)

Für die Computerlinguistik hat sich die Named Entity Recognition inzwischen zu einem der wichtigsten Forschungsgebiete entwickelt. Wer schon einmal den Begriff „Named Entity Recognition“ gehört hat, weiß dass er im Bereich der **Informationsextraktion** (IE) anzusiedeln ist. Im deutschen Sprachraum ist die NER auch unter dem Schlagwort „Eigennamenerkennung“ bekannt.

Doch ist die Erkennung von Eigennamen nicht die einzige Aufgabe der Informationsextraktion, bei der versucht wird, aus Texten nicht-ambige Daten, die ein festgelegtes Format haben, zu extrahieren [Roth, 2002]. Die **Eigennamenerkennung** ist nur eine von verschiedenen Teilaufgaben der IE, wobei sie eigenständig auftreten kann, oder wiederum Teil einer anderen computerlinguistischen Anwendung sein kann. Oft sind Information-Retrieval-Systeme und Systeme zur Antwort-Extraktion, Textzusammenfassung oder maschinelle Übersetzung, sowie Textmining-Programme und Suchmaschinen auf die Dienste der Named Entity Recognition (NER) angewiesen [Roth, 2002].

Leider gehen die Meinungen, was man genau unter Named Entity Recognition zu verstehen hat, auseinander. Der Streitpunkt bei der Definitionsfindung bezieht sich hierbei auf die Klärung des Begriffs der **benannten Entität** (**Named Entity**).

Ohne sich darauf festzulegen, was unter einer benannten Entität verstanden werden soll, lässt sich zunächst folgende Definition geben:

Named Entity Recognition bezeichnet die automatische Erkennung von Instanzen und Einheiten bestimmter Klassen in Texten. Natürlich gibt diese Begriffserklärung keinerlei Aufschluss darüber, welche „Klassen“ von Entitäten nun bei der NER erkannt werden sollen.

Bei [Ciaramita und Altun, 2005] ist der Begriff der *Named Entity* recht eng gefasst, indem sie nur Personen, Organisationen und Orte berücksichtigen, definieren sie:

Named entity recognition (NER) *is the task of tagging words with labels such as person, organization, and location.*

Jedoch werden meist weitere Entitäten, wie Datums-, Zeit-, Prozent- und Währungsangaben mit in die Erkennung von Eigennamen einbezogen.

Named entity recognition (NER) *(...) seeks to locate and classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.* [2]

[2] `http://en.wikipedia.org/wiki/Named_entity_recognition`

Weiterhin bleibt fragwürdig, ob Datums-, Zeit-, Prozent- und Währungsangaben wirklich in die Kategorie der **„benannten Entitäten“** fallen, oder zwar Entitäten, aber keine Eigennamen sind.

Zur Klärung dieser Frage trägt wohl die auf der MUC-7[3] festgelegte Definition zur Erkennung von **„Named Entities“** bei.

The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are 'unique identifiers' of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages).[4]

Auf ein ähnliches Ergebnis kommt man, wenn man vom deutschen Begriff für *Named Entity* – dem „Eigennamen“ – ausgeht.

Denn **Eigennamen** können danach kategorisiert werden, welche Art von Objekt sie bezeichnen: [5]

- Die häufigsten Namensträger sind Personen. Bei Personennamen kann man Vornamen und Familiennamen unterscheiden.
- Eine weitere große Gruppe bilden die Ortsnamen (Toponyme). Diese können weiter untergliedert werden in Städtenamen, Ländernamen, Flussnamen, Flurnamen usw.
- Institutionen sind typischerweise Träger von Eigennamen.
- Eine weitere große Gruppe bilden die Produktnamen.

Als **Eigennamen** werden also Bezeichner für Personen, Orte, Organisationen und Produkte betrachtet. Datums-, Zeit-, Prozent- und Währungsangaben gelten zwar als Entitäten, aber nicht als *Named Entities* – und somit deren Bezeichner auch nicht als Eigennamen [Roth, 2002].

Im Grunde gibt es zwei Arten von NER-Systemen: Die eine Gruppe verwendet linguistische Methoden und die andere baut auf statistischen Modellen auf.

Für den hier vorgestellten Ansatz der automatischen Erkennung von Menschenbezeichnern innerhalb biographischer Relationen in Wirtschaftsnachrichten soll nur die sprachbasierte Named Entity Recognition von Interesse sein.

Hierbei stellen die Personen die wichtigste Entitätsart bei der Eigennamenerkennung dar. Weitere Entitäten, die typischerweise in Wirtschaftstexten auftreten, wie Organisationen, Orte und Zeitangaben werden mit in die Suche einbezogen.

[3] Unter der MUC-7 versteht man die im Jahre 1998 zum 7. Mal durchgeführte **M**essage **U**nderstanding **C**onference.

[4] MUC-7 Named Entity Task Definition [Roth, 2002]

[5] vgl. `http://de.wikipedia.org/wiki/Eigenname`

1.2 Definition: Biographische Relationen

Bis jetzt wurde die Aufgabenstellung der automatischen Erkennung von Menschenbezeichnern innerhalb biographischer Kontexte in Wirtschaftsnachrichten nur in dem Punkt erläutert, welche Entität in den Texten gefunden werden soll.

Inzwischen ist deutlich geworden, dass sich dieser Ansatz auf die Lokalisierung von Eigennamen, insbesondere von Personennamen als „Named Entity“, konzentriert. Dennoch werden nicht nur Namen für Personen betrachtet, sondern auch andere Menschenbezeichnungen, die sich auf den Beruf, die soziale Stellung oder Verwandtschaftsverhältnisse beziehen.

Nun bleibt nur noch zu klären, was unter „biographischen Relationen“ verstanden werden soll. **Biographische Relationen** sind in der Regel Prädikatrelationen von Verben, die vornehmlich in biographischen Kontexten auftreten. Da in biographischen Kontexten die Lebensgeschichte von Personen beschrieben wird, betrifft es Verben, die das Ereignis der Geburt, den schulischen und beruflichen Werdegang, sowie Beziehungen zu anderen Menschen usw. beschreiben.

Diese Prädikate lassen sich aufgrund ihrer Relevanz für die Öffentlichkeit in verschiedene Kategorien unterteilen. Einerseits gibt es eine Gruppe von Verben, die in fast jeder Biographie genannt werden, und andererseits gibt es Verben bzw. Relationen, die nur in Autobiographien zur Sprache kommen. Dies ermöglicht eine Aufspaltung der biographischen Relationen in die drei Unterkategorien der persönlichen, der öffentlichen und der zufälligen Relationen.

1.2.1 Persönliche Relationen

Laut Duden ist eine Biographie nichts anderes als die Niederschrift einer Lebensgeschichte. Somit ist es nicht verwunderlich, wenn manche Leute vieles aus ihrem Leben zu erzählen haben. Dabei werden oft Details aus dem Gefühlsleben preisgegeben, und es wird „aus dem Nähkästchen geplaudert“, wie es in bestimmten Autobiographien der Fall ist. In der Regel werden in diesem Zusammenhang sehr intime Dinge über Personen erzählt, welche für die Öffentlichkeit eigentlich nicht bestimmt sein sollten.

In **persönlichen Relationen** sind besonders solche Verben anzutreffen, die Gefühlsregungen ausdrücken und Informationen aus dem Privatleben der Leute liefern. Doch sind es meist Relationen, die jemanden persönlich betreffen.

Natürlich stellt sich bei diesen Prädikatrelationen nun die Frage, inwiefern sie noch biographische Relevanz haben. An dieser Stelle muss man wohl einräumen, dass persönliche Relationen in ihrer ersten Bedeutung zwar die schönsten Klatschgeschichten aus dem Leben berichten, und somit sicher als biographische Relation gezählt werden können, aber in Wirtschaftsnachrichten kaum Beachtung finden. Deshalb sind sie für den hier vorgestellten Ansatz nahezu irrelevant.

Dennoch ist die Grenze zwischen persönlichen und öffentlichen Relationen manchmal fließend. Ein solcher **„Grenzgänger“** ist meiner Meinung nach das englische Prädikat *„to be married with“*. Diese Beziehung zwischen zwei Leuten wird in den meisten Biographien veröffentlicht. Oft wird noch hinzugefügt, ob es eine glückliche Ehe ist oder war, und wie lange sie schon andauert oder gedauert hat. Die meisten Menschen würden sa-

gen, dass eine Eheschließung ein rechtlicher Akt ist und aufgrund dessen keine Einwände bestehen dürften, dies Außenstehenden mitzuteilen. Doch betrifft eine Ehe immer zwei Personen und ist somit etwas sehr Persönliches. Damit soll nur klar gestellt sein, dass es auch Prädikatrelationen gibt, welche sehr wohl in Lebensläufen öffentlich bekannt gegeben werden dürfen, die dennoch eine starke Brücke zum Privatleben der jeweiligen Personen schlagen.

Im Zuge dieser Arbeit werden nur persönliche Relationen in ihrer zweiten Bedeutung betrachtet. Somit werden nur Prädikatrelationen untersucht, die jemanden persönlich betreffen, wie z.B. *„He was born as son of a blacksmith in 1955.“*.

1.2.2 Öffentliche Relationen

Des Weiteren gibt es eine große Anzahl an Prädikaten, welche in die Kategorie der öffentlichen Relationen fallen. Wie der Name **„öffentliche Relation“** schon verrät, handelt es sich hierbei um Verben, die hauptsächlich in offiziellen Lebensläufen genannt werden und sachliche Informationen aus dem Leben dieser Personen bekannt geben. In der Regel handelt es sich hierbei um Prädikatrelationen, die biographische Fakten beschreiben, welche beispielsweise für die Leser von Wirtschaftsnachrichten von Interesse sein dürften. Hierunter fallen Relationen, welche eventuell Aufschluss darüber geben, welchen Beruf die Person ausübt, oder bei welchem Unternehmen sie gerade beschäftigt ist. Zudem enthalten öffentliche Relationen Details aus dem Leben der jeweiligen Person, bei denen abgeklärt wurde, ob die betreffende Person mit der Veröffentlichung dieser Daten einverstanden war. Manchmal ist dies auch nicht der Fall, wie der Artikel *„Google balances privacy, reach“* von [Mills, 2005] (siehe Vorwort) gezeigt hat. Doch die Missachtung der Privatsphäre bei der Informationssuche ist ein anderes Thema und macht die gefundenen Fakten nicht weniger offiziell.

Da die Aufgabe der Erkennung von Menschenbezeichnern innerhalb biographischer Relationen sich auf die Domäne der englischsprachigen Wirtschaftsnachrichten beschränken wird, werden öffentliche Relationen im Zentrum dieser Untersuchung stehen.

1.2.3 Zufällige Relationen

Vollständigkeitshalber sollten auch die **„zufälligen Relationen“** angesprochen werden. Denn im Leben der Menschen gibt es enorm viele zufällige Begebenheiten. Auch über sie lassen sich zahlreiche Geschichten erzählen. Oft können zufällige Ereignisse zusammen mit persönlichen Gefühlen auftreten und dann vermischen sich wieder persönliche mit zufälligen Relationen. Leider sind zufällige Prädikatrelationen am uninteressantesten für das Auffinden von Personen in biographischen Kontexten, da sie kaum vorhersehbar sind und in einer solchen Vielfalt vorkommen, dass sie schwer aufzuzählen sind. Außerdem wird einem Beinbruch, einer Verliebtheit oder einem Streit in der Familie meist wenig Beachtung von Außenstehenden geschenkt.

Biographien schreibt das Leben –
welche Art von Information bzw. Relation sie enthalten, hängt allein vom Autor ab.

1.3 Einschätzung der Thematik

1.3.1 Probleme und Chancen

Die Aufgabe der automatischen Erkennung von Menschenbezeichnern innerhalb biographischer Relationen in englischsprachigen Wirtschaftsnachrichten wird sicher kein leichtes Unterfangen werden. Doch ist es eine Herausforderung, der man sich ohne Weiteres stellen kann, indem man sich zunächst ein Bild von der Ausgangssituation macht und sich danach die möglichen Schwierigkeiten vor Augen führt.

Einerseits ist es wichtig, vorab abzuklären, welche Entitäten, Bezeichnungen oder andere Angaben in biographischen Kontexten vorkommen.

So kann ein Personenname beispielsweise aus einem Titel oder einer Anrede gefolgt von einem Nachnamen bestehen. Des Weiteren werden in biographischen Texten häufig Beschäftigungsverhältnisse beschrieben, und in diesem Zusammenhang werden sicherlich Organisationsnamen bzw. Firmennamen auftreten. Auch eine Liste der Branchen, Fachbereiche und Industriesektoren kann von Vorteil sein, wenn nur die Arbeitsdomäne einer Person genannt wird. Zudem kommen in diesen Kontexten häufig Ortsbestimmungen und Beschäftigungszeiträume vor.

Weiterhin wäre es sinnvoll, Wörterbücher für Titel und Anredemöglichkeiten zu erstellen, sowie Vor- und Nachnamen aufzulisten, aber auch vollständige Personennamen zu archivieren. Außerdem lassen sich weitere benannte Entitäten wie Toponyme und Organisationen ebenfalls mit der Hilfe von Lexika in den Griff bekommen. Mit anderen Kategorien von Bezeichnern kann ähnlich verfahren werden, so dass Hyperonymierelationen in Form von Wörterbüchern kodiert und somit semantische Klassen gebildet werden.

Andere linguistische Phänomene lassen sich dagegen schlecht mittels Lexika beschreiben, dafür können sie gut über lokale Grammatiken dargestellt werden.

- Darunter fällt z.B. die **syntaktische Variabilität**. Gerade wenn man an die Beschreibung von Datumsangaben oder Personennamen denkt, gibt es eine Reihe an syntaktischen Möglichkeiten, wie diese ausgedrückt werden können.

```
on February 20, 2004
on 3 June 1994
on Tuesday 6th April 2005
12-Feb-2006
in March 1960

Bill Gates
Mr. Gates
William Henry Gates III
William Gates
```

- Ein weiteres klassisches Problem ist die Unterscheidung zwischen einer Firma, ihrer Marke und ihrem Produkt. Dafür wäre *„Apple“* ein Paradebeispiel, denn

allein der Kontext, in dem dieser Begriff fällt, könnte für die Auflösung dieser Ambiguität, sorgen. Solche **Disambiguierungen** können mit der Hilfe von lokalen Grammatiken relativ leicht und zugleich recht anschaulich vorgenommen werden.

- Zudem tragen sie nicht nur zur Bedeutungsunterscheidung innerhalb von *Named Entities* bei, sondern auch zwischen Eigennamen und allgemeinen Bezeichnungen. Beispielsweise gibt es einige Nachnamen, welche gleichzeitig auch in ihrer Funktion als Nomen eine Pflanze wie *„Bush“*, eine Berufsbezeichnung wie *„Miller“* oder eine Farbe wie *„Blue“* sein können. Nur eine detaillierte Beschreibung des Kontextes durch eine lokale Grammatik kann verhindern, dass z.B. *„The Burning Bush“* keine Person bezeichnet, aber *„Bush jr.“* auf jeden Fall einen Menschen benennt.

1.3.2 Bewältigung der Aufgabe

Wie aus dem letzten Abschnitt hervorgeht, gibt es einiges zu bedenken, wenn man sich an die Aufgabe heranwagt, eine Grammatik für Menschenbezeichner in biographischen Kontexten zu entwickeln.

Aufgrund dessen werden die folgenden Kapitel einen Einblick in die Herangehensweise an dieses Thema geben und dabei die entsprechenden Lösungsansätze präsentieren.

Zunächst werden Begrifflichkeiten, Bedeutung und Funktionalität von lokalen Grammatiken in Kapitel 2 erläutert. Im Anschluss daran wird noch im selben Kapitel auf die Arbeitsweise mit dem System UNITEX eingegangen, um deutlich zu machen, wie mit lokalen Grammatiken gearbeitet werden kann.

Nachdem die Grundlagen zu Grammatiken gelegt wurden, können in Kapitel 3 interessante Ansätze weiterer Linguisten vorgestellt werden, die große Fortschritte auf dem Gebiet der automatischen Erkennung von Eigennamen mittels lokaler Grammatiken erzielt haben und deren Arbeiten meinen Ansatz zur Erkennung von Menschenbezeichnern in biographischen Kontexten geprägt haben.

In Kapitel 4 werden alle gewollten Beschränkungen für meinen Ansatz zur Erkennung von Menschenbezeichnern in biographischen Kontexten beschrieben. Es werden Erklärungen gegeben, warum man sich auf die Korpusdomäne der Wirtschaftsnachrichten festgelegt hat, und wieso die Personen gegenüber anderer Entitäten im Vordergrund stehen.

Daraufhin werden Einzelheiten zu den im System verwendeten Ressourcen preisgegeben. Dabei werden in Kapitel 5 die verschiedenen Korpora und alle selbst erstellten Lexika angesprochen, mit deren Hilfe dieser Ansatz zur Erkennung von Eigennamen umgesetzt wurde.

Die folgenden Kapitel stellen die entwickelten lokalen Grammatiken für Entitäten – wie Personen, Organisationen, Toponyme und Datumsangaben – vor und geben Einblick in die Grammatiken der persönlichen, sowie beruflichen Relationen.

Zuletzt wird die Qualität des Systems gemessen, indem die Ergebnisse der lokalen Grammatiken auf einem Testkorpus evaluiert werden. Außerdem wird aufgezeigt, wie persönliche und berufliche Relationen aus dem Text extrahiert werden können, um die syntaktische und semantische Vielfalt dieser Prädikatrelationen zu veranschaulichen.

2 Lokale Grammatiken

2.1 Was sind lokale Grammatiken?

Lokale Grammatiken kann man als „Landkarten der Sprache“ bezeichnen [Mallchok, 2004], die einerseits Sequenzen von Wörtern, welche semantische Einheiten bilden, und andererseits syntaktische Strukturen beschreiben.

Überdies geben sie noch Aufschluss über die morphosyntaktischen Eigenschaften, der darin beschriebenen Elemente, welche syntaktisch [Fairon, 2000] oder semantisch [Constant, 2000] geprägt sein können.

Des Weiteren können sie in den verschiedensten Varianten für automatische Sprachverarbeitung auf Textkorpora nützlich sein. Besonders auf dem Gebiet der lexikalischen Disambiguierung werden lokale Grammatiken verstärkt eingesetzt [Blanc und Dister, 2004].

Da Wortformen isoliert gesehen oft ambig sind, kann ein Teil von ihnen aber durch die Analyse des Kontextes disambiguiert werden. Der für die Disambiguierung relevante Kontext wird durch eine lokale Grammatik [Gross, 1997] beschrieben, die durch einen endlichen Automaten bzw. einen Transduktor repräsentiert wird. Lokale Grammatiken werden nicht nur für die Disambiguierung, sondern auch für andere Aufgaben genutzt, wie die Erkennung von Mehrwortlexemen und Komposita, die Repräsentation orthographischer Varianten im Lexikon, sowie die Überprüfung der Kongruenz oder Identifikation von Zeitangaben und anderen Entitäten [Blank, 1997].

Endliche Automaten bzw. Transduktoren beschreiben komplexe linguistische Strukturen, die so nicht in einer Lexikongrammatik oder in elektronischen Wörterbüchern formalisiert werden könnten. Eigentlich sind Transduktoren endliche Automaten, die zusätzlich eine Ausgabe erzeugen, wenn die in der Definition des Automaten spezifizierte(n) Sequenz(en) erkannt wurde(n). Der „Eingabeteil“ des Transduktors dient dazu, spezifische Sequenzen im Text zu erkennen. Der „Ausgabeteil“ führt einerseits Substitutionen im Text aus, versieht andererseits identifizierte Sequenzen mit zusätzlichen Informationen (z.B. einer Wortklasse) oder fügt linguistische Markierungen (z.B. die Annotation von Phrasen) in den Text ein [Blank, 1997].

In der Regel werden lokale Grammatiken in Form von Graphen [Silberztein, 1993] visualisiert. Die Kombination von parametrisierten Graphen mit einer Lexikongrammatik kann beispielsweise äußerst effektiv bei der syntaktischen Analyse einfacher Sätze sein [Paumier, 2001; Laporte, 2005].

Graphen sind sehr geeignete Repräsentationen für lokale Grammatiken, denn es gibt diverse Grafikprogramme, mit denen sich diese Graphen leicht erstellen, erweitern oder abändern lassen. Die beiden Systeme INTEX und UNITEX bieten u.a. solche Zeichenprogramme für Automaten an.

Jeder Graph besteht aus einem Anfangszustand, der durch einen Rechtspfeil symbolisiert wird. Dieser Rechtspfeil geht von keinem Zustand aus, sondern führt lediglich zu einem der nächsten Zustände im Graphen. Außerdem enthält jeder endliche Graph einen Endzustand, welcher meist durch einen doppelt umrandeten Kreis dargestellt wird. Die Graphen werden von links nach rechts interpretiert und so werden die möglichen Pfade „abgelaufen“ und ihre Muster im Text gesucht. Bei den Systemen INTEX und UNITEX steht jeder Zustand bzw. jeder Knoten für Wörter (mit oder ohne ihrer morphologischen Informationen) oder für Klassen aller Flexionsformen von Wörtern, wenn diese in spitzen Klammern notiert wurden. Somit werden die Eingabesequenzen des Transduktors nicht an den Übergängen zu den Zuständen genannt, sondern in den Zuständen selbst. Natürlich sind auch wie bei endlichen Automaten ϵ-Transitionen erlaubt. Alle Transitionen werden durch Verbindungslinien zwischen den einzelnen Zuständen dargestellt. Das leere Wort wird als `<E>` in den Knoten angegeben. Es wird sogar gestattet Subgraphen innerhalb eines Automaten aufzurufen, was die Übersichtlichkeit der Graphen erhöht. Diese Subgraphen werden grau unterlegt, so dass eine Unterscheidung zwischen einem einfachen Zustand und einem Zustand, der einen weiteren Graphen aufruft, möglich wird.

Die eben beschriebenen Graphen sind auch als gerichtete azyklische Graphen bekannt, da sie keinerlei Zyklen enthalten. Im englischen Sprachraum werden sie als *„Directed Acyclic Graphs“* bezeichnet und werden deshalb im deutschen Sprachraum häufig nur DAGs genannt. Mathematisch gesehen repräsentiert ein DAG eine Halbordnung.

2.2 Warum werden lokale Grammatiken verwendet?

Die meisten Versuche linguistische Theorien oder Grammatiken zu entwickeln, welche umfassend und stark verallgemeinert beschreiben wollen, wie eine Sprache aufgebaut ist, und wie Syntax, Morphologie und Semantik zusammenwirken, waren wenig befriedigend. Denn Ziel solch einer Grammatik sollte es immer sein, alle Sätze, die in einer Sprache möglich sind, abzudecken, und kein Satz, der mit dieser Grammatik gebildet werden konnte, durfte grammatikalisch oder semantisch unstimmig sein.

Anfangs ging man an dieses Problem so heran, dass jede explizite Komponente im Satz durch ihre jeweilige grammatikalische Kategorie ersetzt wurde. Noam Chomsky fasste 1957 diese Grammatiken unter dem Begriff *„Kontextfreie Grammatik“* zusammen, musste aber einräumen, dass es immer noch einige Unzulänglichkeiten in Bezug auf die formale Repräsentation natürlicher Sprache gab. Diese Grammatiken beschrieben in der Regel nur einfache Sätze und gingen kaum auf die Abhängigkeiten der einzelnen Satzteile untereinander ein [Gross, 1997].

Dagegen waren die späteren Ansätze von Zellig Sabbetai Harris und Noam Chomsky schon spezieller, da sie inzwischen Bildungsregeln für die einfachen Sätze definierten und

diese dann untereinander kombiniert wurden, so dass komplexe Sätze geformt werden konnten. Im Grunde war es damals schon ein kleiner Schritt in Richtung Diskursanalyse, den die beiden vollzogen. Denn sie legten Regeln fest, welche die Satzstellung innerhalb der einfachen Sätze variierten und einfache Sätze zu komplexen Satzgefügen verbanden.

Irgendwann stellte sich dann heraus, dass diese theoretische Sichtweise der natürlichen Sprache, die immer komplexer werdenden Beschreibungsformalismen und die vielen Ausnahmen, welche sich in die Bildungsregeln eingeschlichen hatten, nicht mehr zu handhaben waren. Daraufhin besonnen sich viele Linguisten darauf das Phänomen „Sprache“ anders zu erforschen. In ihrer Herangehensweise verhielten sie sich ähnlich wie Naturwissenschaftler. Man muss keine Sätze erfinden, denn es gibt sie schon, und man muss das Vorhandene zuerst untersuchen, bevor Neues automatisch generiert werden kann. Laut Maurice Gross findet man eine Grammatik im Text und muss sich nicht erst eine ausdenken.

Deshalb sollte man als Linguist keine Theorie in die Welt setzen, bevor man nicht Korpusmaterial gesammelt und seinen Ansatz auf realem Text verifiziert hat. Denn indem Satzkorpora gebildet werden, deren syntaktische und semantische Struktur analysiert wird, entstehen indirekt schon Regeln zur Beschreibung der Sprache.

Des Weiteren war Zellig S. Harris davon überzeugt, dass die Untersuchung von Subsprachen in Verbindung mit lokalen Grammatiken besonders vielversprechend sein dürfte, weil Subsprachen

- thematisch begrenzt sind,
- lexikalischen, syntaktischen und semantischen Restriktionen unterliegen,
- in ihren grammatikalischen Eigenschaften nicht der Allgemeinsprache gleichen,
- gewisse lexikalische Strukturen relativ häufig wiederholen
- in sich strukturiert sind und
- eine gewisse Symbolik verwenden.

So können Elemente der Sprache, die in lokalen Grammatiken erfasst werden, als kleine, aber aussagekräftige Subsprachen gesehen werden, und Beschreibungsversuche von Subsprachen würden in ihrer Repräsentation erweiterten lokalen Grammatiken entsprechen.

Die Einschränkung der Sprache auf eine bestimmte Bezugsdomäne – wie z.B. auf Wirtschaftsnachrichten – und die damit verbundene Verwendung von themenspezifischen Fachvokabular rechtfertigen gewiss den Einsatz von lokalen Grammatiken. Aufgrund dessen sind lokale Grammatiken zur syntaktischen und semantischen Analyse von Menschenbezeichnern innerhalb biographischer Relationen sicherlich die richtige Entscheidung.

2.3 UNITEX – Ein System zur Anwendung lokaler Grammatiken

UNITEX ist ein Korpusverarbeitungssystem, welches es ermöglicht, mit elektronischen Ressourcen wie z.B. elektronischen Lexika umzugehen und lokale Grammatiken zu entwickeln und anzuwenden. Dabei wird auf drei Ebenen der Sprache – der Morphologie, dem Lexikon und der Syntax – gearbeitet.

Die Hauptfunktionen von UNITEX sind u.a

- das Erzeugen, sowie die Anwendung und Verarbeitung elektronischer Wörterbücher,
- die Benutzung von regulären Ausdrücken zum Pattern Matching,
- die Interpretation rekursiver Transitionsnetze zum Pattern Matching,
- die Anwendung von lokalen Grammatiken und Lexikongrammatiken und
- die Auflösung von Ambiguitäten über den Text-Automaten.

Das Konzept für das System UNITEX wurde am LADL (*Laboratoire d'Automatique Documentaire und Linguistique*) unter der Leitung von Prof. Maurice Gross entwickelt, und das dazugehörige Programm wurde am Institut Gaspard-Monge (IGM) der Université de Marne-la-Vallée von Sébastien Paumier implementiert.

Derzeit werden für UNITEX Lexika in 14 verschiedene Sprachen (Deutsch, Englisch, Finnisch, Französisch, Griechisch, Italienisch, Koreanisch, Norwegisch, Polnisch, Portugiesisch, Brasilianisches Portugiesisch, Russisch, Spanisch und Thai) angeboten.

Da UNITEX im Gegensatz zu INTEX frei verfügbar ist und unter der GNU GPL (*GNU General Public License*) steht, kann es im Grunde jeder benutzen. Außerdem stellt es ganz ähnliche Funktionen wie INTEX zur Verfügung und ist auf allen gängigen Betriebssystemen (Windows, Linux, MacOS) lauffähig.[6]

Vorallem bietet UNITEX eine komfortable und intuitiv bedienbare Oberfläche zur Entwicklung von Grammatiken. Dabei handelt es sich um eine Java-Oberfläche, von der aus diverse C++-Programme gesteuert werden.

2.3.1 Textvorverarbeitung

UNITEX arbeitet mit der Kodierung „UTF-16 Little Endian“ und unterstützt somit den Unicode 3.0 Standard. Dadurch wird selbst die Verarbeitung asiatischer Sprachen ermöglicht. Zur Konvertierung der Texte empfiehlt sich das Programm *Convert* von UNITEX. Nachdem UNITEX mit der gewählten Sprache gestartet worden ist, kann man einen Text mit der Kodierung UTF-16 LE öffnen. Dabei wird gefragt, wie der Text vorverarbeitet werden soll. Die Textvorverarbeitung von UNITEX setzt sich aus den Schritten Normalisierung, Satzenderkennung, Auflösung von Kontraktionen, Tokenisierung und lexikalische Analyse des Korpus zusammen.

[6] `http://www-igm.univ-mlv.fr/~unitex/download.html`

Normalisierung

Es ist Aufgabe des Programms *Normalize* die Normalisierung des Textes vorzunehmen, indem Folgen von Leerzeichen bzw. Zeilenumbrüchen durch ein Zeichen ersetzt werden. Gleichzeitig wird die interne Syntax von eventuell lexikalisch annotierten Token überprüft.

Satzenderkennung und Auflösung von Kontraktionen

UNITEX bietet eine sprachspezifische Satzenderkennung mittels lokaler Grammatiken in Form von Graphen an. Des Weiteren werden Kontraktionen wie z.B. *„I'm“* zu *„I am“* oder *„you're“* zu *„you are“* aufgelöst und verschiedene Arten von Anführungszeichen vereinheitlicht.

Tokenisierung

Hierfür ist das Programm *Tokenize* von UNITEX zuständig. Die Tokenisierung wird aufgrund des Alphabets der jeweiligen Sprache vorgenommen. Die daraus resultierende Tokenliste wird für spätere Zwecke im Arbeitsverzeichnis des aktuellen Textes gespeichert.

Lexikalische Analyse

Bei der lexikalischen Analyse werden alle Standardwörterbücher der jeweiligen Sprache und eventuell noch eigene Lexika auf die Tokenliste angewendet. Dabei kommt das Programm *Dico* zum Einsatz, welches alle Token mit der entsprechenden grammatikalischen oder semantischen Information aus den Lexika versieht. Alle Lexika, welche vom System UNITEX verwendet werden sollen, müssen formal dem Standard der DELA Wörterbücher entsprechen.

2.3.2 DELA Wörterbücher [Geierhos, 2005]

Das klassische Wörterbuch ist eine Sammlung von Wörtern oder einer Kategorie von Wörtern einer Sprache, die in der Regel in alphabetischer Ordnung mit Erläuterungen in derselben Sprache oder einer Übersetzung derer in eine andere Sprache aufgelistet sind (Lexis, 1975). Dagegen ist das elektronische Wörterbuch eine formale Repräsentation eines Lexikons, welche jeder Flexionsform ihr Lemma, genauso wie die entsprechende grammatikalische, Flexions- und eventuelle semantische Information zuweist (nach Sébastien Paumier)[7].

Überdies hinaus wird von einem elektronischen Wörterbuch gefordert, dass es formal und vollständig ist, so dass es sich maschinell verarbeiten lässt und es von Programmen automatisch verändert werden kann. Theoretisch müsste es 100% des Lexikons abdecken, was allerdings kaum realisierbar ist.

[7] Übersetzung aus dem Französischen
http://wwwigm.univmlv.fr/~paumier/DEA/Cours%206%20%20Dictionnaires%20electroniques.pdf

DELA ist ein elektronisches Wörterbuchsystem und steht für *„Dictionnaires électroniques du LADL“*[8]. In den 60er Jahren wurde es von Prof. Maurice Gross ins Leben gerufen, und war zunächst unter dem Namen „Lexikon Grammatik“ bekannt. Das DELA ist eine formale Repräsentation der jeweiligen Sprache; das heißt, Spracheigenschaften werden strukturiert abgespeichert, wobei sowohl Vokabular als auch Morphologie berücksichtigt werden.

Die DELA-Wörterbuchfamilie gliedert sich in folgende Teillexika:

- DELAS *„mots simples“*: Wörterbuch für die einfachen Wörter
- DELAC *„mots composés“*: Wörterbuch für die komplexen Wörter
- DELAF *„formes fléchies“*: Wörterbuch der einfachen Wörter, deren Flexionsmerkmale kodiert sind.
- DELACF *„mots composés avec les formes fléchies“*: Wörterbuch der komplexen Wörter, deren Flexionsmerkmale kodiert sind.

Dabei werden als einfache Wörter (*„mots simples“*) Sequenzen zusammenhängender Buchstaben eines Alphabets einer bestimmten Sprache verstanden, wie z.B. `angry,.A` oder `acually,.ADV` oder `bodies,body.N:p`.

Dagegen sind komplexe Wörter (*„mots composés“*) Sequenzen zusammengesetzter lexikalischer Einheiten wie einfache Wörter, Trennzeichen oder Ziffern.

Beispiele aus dem Französischen wären hierfür `coup de chance,.N+NDN:ms` (Glückstreffer) oder `coup de pied,.N+NDN:ms` (Fußtritt) oder das ambige `coup de foudre,.N+NDN:ms` (Liebe auf den ersten Blick /Blitzschlag).

Die eben genannten Beispiele deuteten bereits an, dass hinter einem Eintrag im DELAF eine gewisse Symbolik steht.

So besteht ein Lexikoneintrag im DELAF aus 5 verschiedenen Feldern [Courtois, 2004]:

1. Flektierte Form des Wortes
2. Lemma des Wortes (Kanonische Form)
3. Charakteristische Informationen zur Lemmaform
4. Grammatikalische Eigenschaften der flektierten Form
5. Optionale Ergänzungen für den menschlichen Betrachter

Analog dazu wird ein Eintrag im DELACF gebildet. Dabei sollte man noch anmerken, dass das zweite Feld (die Lemmaform) immer dann leer ist, wenn sie mit der flektierten Form identisch ist. Dafür wird das vierte Feld (die grammatikalische Information für die flektierte Form) nicht belegt, wenn das Wort eindeutig ist, und es nicht variiert

[8] LADL = Laboratoire d'Automatique Documentaire et Linguistique

werden kann. Außerdem wird das fünfte und letzte Feld (die Zusatzinformation) nur besetzt, wenn die flektierte Form – das Ausgangswort – ein Kompositum ist. Genau die gleichen Regeln gelten für Lexikoneinträge im DELAS und DELAC, nur dass hier die Flexionsinformation entfällt.

An einem konkreten Beispiel würde dies nun folgendes bedeuten:

```
bodies,body.N:p
```

⇓

```
bodies : flektierte Form
body   : Lemmaform
N      : grammatikalische Information (Nomen)
p      : grammatikalische Eigenschaft der flektierten Form (Plural)
```

Bei der Erstellung eigener Lexika sollte darauf geachtet werden, dass Mehrwortlexeme direkt im Lexikon kodiert werden, weil sonst Fehler bei der Tokenisierung gemacht werden. Wenn man nur ein Teilformenlexikon verwenden würde, könnte beispielsweise *„grand-mère“* nicht als ein Wort erkannt werden. Oft besteht auch die Möglichkeit Mehrwortlexeme wie *„grand-mère“* (Großmutter) anstatt des Bindestrichs mit einem Leerzeichen dazwischen zu schreiben. Dafür wäre dann `grand=mères,grand=mère.N:fp` der entsprechende Lexikoneintrag, denn das `'='` ist ein Metazeichen, was für einen Bindestrich `'-'` und für ein Leerzeichen `' '` steht.

Je nachdem wie ausführlich die Kodierung eines Lexikons mit diversen grammatikalischen oder semantischen Angaben vorgenommen wurde, spricht man von 3 Stufen der Lexikonkodierung:

- **DELAF-S** (*„short“*): Es werden minimale Angaben zur grammatikalischen Analyse der einzelnen Formen gemacht. Das heißt, dass lediglich Informationen zur jeweiligen Wortart und zur Flexion kodiert werden. Hier wird ausschließlich auf die *Grammatik* Bezug genommen.

- **DELAF-M** (*„medium“*): Die Lexikoneinträge werden um semantische Informationen zu den Nomina erweitert. Dabei wird spezifiziert, welche Eigenschaften das Nomen hat, z.B. ob es ein Menschenbezeichner `Hum`, ein Konkreta `Conc` oder ein Tier `Anl` etc. ist. Außerdem werden Determinativa `DET` und Pronomina `PRO` durch weitere Unterkategorien versehen. Auf diese Weise wird die *Semantik* miteinbezogen.

- **DELAF-L** (*„large“*): Hierbei werden die Wörterbucheinträge um die Lexikongrammatik der LADL ergänzt, so dass die syntaktischen Eigenschaften der Verben im Französischen markiert werden (Berücksichtigung der *Syntax*).

Wie ausführlich nun ein Lexikoneintrag erstellt wird, hängt ganz von seiner späteren Funktion ab und über welche Art von Informationen er später angesprochen werden soll. Das heißt nichts anderes, als dass beispielsweise Nomina, welche die semantische Funktion eines Menschenbezeichners haben, auch als solche markiert werden sollten. Legt man allerdings nachher Wert auf Kongruenzeigenschaften, so sollte man auf keinen Fall die grammatikalische Information außer Acht lassen.

2.3.3 Prioritäten bei der Anwendung der Lexika

UNITEX unterscheidet drei Prioritäten bei der Anwendung der Lexika, falls der Dateiname eines Lexikons (ohne die Endung .bin) auf `'-'` bzw. `'+'` endet:

1. *-.bin (höchste Priorität – diese Lexika werden vorrangig behandelt)
2. *.bin (durchschnittliche Priorität – diese Lexika werden zweitrangig behandelt)
3. *+.bin (niedrigste Priorität – diese Lexika werden zuletzt auf den Text angewendet)

Token, die einem der Lexika einer Prioritätsebene gefunden wurden, werden in keinem Lexikon mit untergeordneter Priorität mehr nachgeschlagen. So lassen sich z.B. bestimmte Lesarten für ein Token erzwingen, da das höher priorisierte Lexikon wie ein Filter andere Bedeutungen aussiebt. Innerhalb einer Prioritätsebene werden alle Lexika gleichrangig behandelt, d.h. verschiedene Lesarten eines Tokens aus unterschiedlichen Lexika werden ins Textlexikon geschrieben.[9]

2.3.4 Mustererkennung und Konkordanzen

Wie bereits erwähnt, werden lokale Grammatiken im System UNITEX als Graphen (DAGs) repräsentiert. Möchte man nun eine lokale Grammatik auf einem Korpus testen, so wählt man den entsprechenden Graphen aus, und das Programm *Locate* wendet diesen Graphen auf den Text an und erstellt den Index für eine Konkordanz. Dabei bietet *Locate* dem Benutzer verschiedene Arten der Textsuche an, bei der

- die kürzesten Treffer,
- die längsten Treffer oder
- alle Treffer

ausgegeben werden.

Außerdem lässt sich das Verhalten des Graphen steuern, falls es sich um einen Transduktor handelt. Es gibt folgende Möglichkeiten:

- Die Ausgabe des Transduktors bleibt unberücksichtigt.
- Die Ausgabe des Transduktors wird links vom Treffer eingefügt.
- Die gefundene Sequenz wird durch die Ausgabe des Transduktors ersetzt.

Für das Anfertigen einer Konkordanz ist das Programm *Concord* zuständig. Es gibt einerseits die Konkordanz in verschiedenen Formaten aus (HTML, Text), und andererseits lässt sich die Länge des Kontextes und die Sortierweise der Treffer spezifizieren.

[9] vgl. `http://www.cis.uni-muenchen.de/~wastl/lg/introUnitex.pdf`

3 Zusammenfassung früherer Arbeiten

In diesem Kapitel werden einige für diesen Ansatz relevante Arbeiten vorgestellt, welche interessante Methoden bei der automatischen Erkennung von Eigennamen, aber auch allgemeine praktische Hinweise im Umgang mit lokalen Grammatiken beschreiben.

3.1 Bootstrapping

3.1.1 Bootstrapping bei der Entwicklung lokaler Grammatiken [Gross, 1999]

Bereits in Kapitel 2 wurden einzelne Aufsätze von Maurice Gross, welche das Konzept hinter den lokalen Grammatiken erläutern, zitiert. Dabei wurde ein weiterer Artikel, der besonders für die praktische Arbeit mit lokalen Grammatiken wichtig ist, außer Acht gelassen. *„A Bootstrap Method for Constructing Local Grammars“* [Gross, 1999] sollte deshalb in diesem Zusammenhang nicht ungenannt bleiben.

Maurice Gross stellt hierbei einen Ansatz vor, lokale Grammatiken oder elektronische Lexika um ein Wort oder Mehrwortlexem herum zu entwickeln.

Dabei wird zunächst vom vorhandenen Lexikoninventar ausgegangen und jeder Eintrag als solch ein Schlüsselbegriff gesehen. Mithilfe einer Suchfunktion kann dann der jeweilige Kontext zu den Ausgangsbegriffen auf einem Beispieltext ermittelt werden. Zu jedem neuen Vorkommen im Korpus wird der Kontext entsprechend seiner lexikalischen Funktion manuell ausgewertet. So können relativ schnell und auf einfache Art Mehrwortlexeme gefunden werden, die das Schlüsselwort in irgendeiner Form enthalten. Als nächstes empfiehlt es sich, die unmittelbaren Kontexte des Ausgangswortes zu schematisieren, um später gezielt alle möglichen Äußerungen, welche diesen Begriff enthalten, abzudecken.

Im konkreten Fall heißt das, dass beispielsweise das Wort *„health“* der Schlüsselbegriff war, und in der Konkordanz lässt sich das Muster `health and <N>` identifizieren. Natürlich werden noch viele andere linguistische Schemata erkennbar sein, doch geht man jedes einzeln durch. Ausgehend von diesem regulären Ausdruck kann man eine Grammatik in Form eines Finite-State-Graphen entwerfen, mit der alle Nomina gefunden werden, welche in diesem speziellen Kontext von *„health“* auftreten.

Daraus ergibt sich dann der Graph aus Abbildung 3.1. Dieser lässt sich erneut in einen anderen Graphen einbinden, welcher den linken Kontext von *„health“* spezifiziert. Diese Methode kann nun beliebig oft wiederholt werden, bis alle möglichen Kontexte des Ausgangswortes ermittelt und beschrieben wurden.

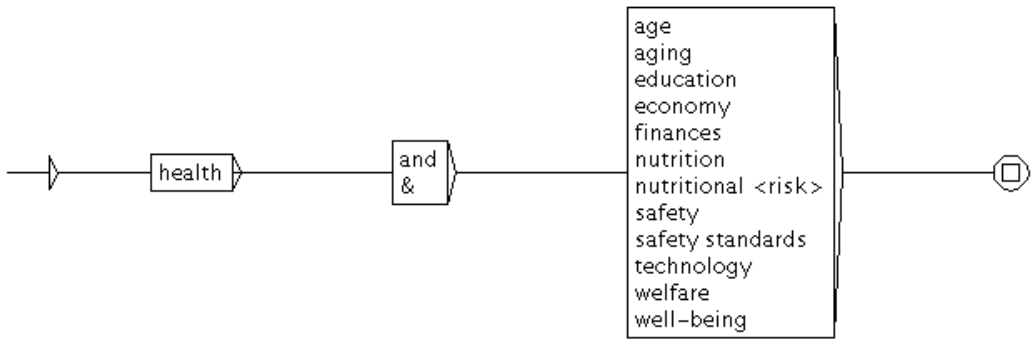


Abbildung 3.1: *HealthAndN.grf* aus [Gross, 1999]

Diese Methode zur Entwicklung lokaler Grammatiken wird als **Bootstrapping** bezeichnet und liefert systematisch und schnell gute Ergebnisse bei der Grammatikerstellung.

3.1.2 Bootstrapping zur Erkennung von Nominalphrasen mit FSTs [Senellart, 1998b]

Jean Senellart stellt in seinem Aufsatz *„Tools for locating noun phrases with finite state transducers“* [Senellart, 1998b] verschiedene praxisnahe Verfahren (Tools) vor, wie man relativ schnell eine große Datenbasis endlicher Automaten (FSTs[10]) aufbauen kann, welche Eigennamen und Berufsbezeichner in Nominalphrasen lokalisieren.

Die Vorgehensweise

Anfangs wird nur ein Wort ausgewählt, zu dem die Konkordanz über den Text erstellt wird. In diesem Fall war es das Wort *„officer“*, welches als Ausgangspunkt für die Graphenkonstruktion diente. Mithilfe der Konkordanzen konnte man unter anderem feststellen, welche militärischen Ränge im Korpus zusammen mit *officer* vorkamen und so Subgraphen erstellen, welche dies abdeckten. Des Weiteren traten auch Adjektive und Nomen im Kontext von *officer* auf, welche Staatszugehörigkeiten ausdrücken. Um diese Ergänzungsmöglichkeiten bzw. Spezifikationen nicht zu verlieren, entschied man sich sie in Form von Wörterbüchern zu kodieren. Auf diese Weise können die gewonnenen Erkenntnisse vielschichtig eingesetzt werden und sind nicht nur an diesen Kontext gebunden. Gleiches gilt für die gesammelten Kontexte von *officer*, denn auch diese konnten bei anderen Berufsbezeichnungen mit Erfolg angewendet werden.

Nachdem man Graphen zu einem bestimmten Schlüsselwort erstellt hat, ist es auch möglich diese Graphen zu verwenden, um neue Begriffe zu finden, die den gleichen Kontext wie das Schlüsselwort aufweisen. Dafür muss lediglich der ursprüngliche Schlüsselbegriff durch eine Variable ersetzt werden. Der neue Graph liefert dann beim Matching

[10]FSTs = Finite-State-Transducers. Sie können sowohl als Transduktoren, aber auch als endliche Automaten fungieren. Ob eine Ausgabesequenz erzeugt werden soll, hängt allein vom Benutzer ab.

alle Ergebnisse, die der alte Graph gefunden hat, und neue Treffer, die nun an der Stelle von *officer* andere Berufsbezeichner aufweisen. Mit dieser Methode lassen sich leicht Synonyme, Hyponyme oder andere semantisch ähnliche Wörter zum Ausgangsbegriff finden.

Mithilfe dieser neuen Kandidaten können nun die beiden eben genannten Schritte wiederholt werden, und so entsteht eine Dynamik in der Grammatikentwicklung, bei der aus alten Ergebnissen immer neuere, bessere und ausbaufähigere Resultate entstehen. Genau diesen dynamischen Entstehungsprozess versteht man hier als **Bootstrapping**.

Das Ergebnis

Insgesamt hat Jean Senellart mehr als 200 verschiedene Graphen konstruiert, um so viele Berufsbezeichnungen wie möglich abzudecken, und seine Lexika zu Nachnamen und Städten enthielten jedes für sich einige tausend Einträge. Seiner Meinung nach war das auch ein praktischer Beweis dafür, dass die Entwicklung lokaler Grammatiken sehr effizient sein kann, wenn diese nahe am Text verläuft. Doch diese Effizienz wurde auch durch die entsprechende Software gefördert, denn ein Graphen-Editor, ein Index basierter Parsing-Algorithmus, sowie ein Konkordanzprogramm und diverse Debugging-Möglichkeiten sind Tools, die bei der Konstruktion von FSTs sehr hilfreich sein können.

3.2 Lemmatisierung zusammengesetzter Zeiten im Englischen [Gross, 1998-1999]

Für die Erkennung von Menschenbezeichnern in biographischen Kontexten, kann es durchaus hilfreich sein, alle Verbkonstruktionen zu lokalisieren, da diese häufig die Semantik des biographischen Ereignisses tragen. Auf diese Weise können Verbkonstruktionen innerhalb der zu untersuchenden Sätze von den potentiellen Entitäten (z.B. Personennamen) in Subjekt- oder Objektposition abgegrenzt werden.

Zu diesem Zweck hat Maurice Gross in den Jahren 1998/1999 ein sehr umfangreiches Graphenpaket zur Lemmatisierung zusammengesetzter Zeiten im Englischen entwickelt, welches er in seinem Aufsatz *„Lemmatization of compound tenses in English“* [Gross, 1998–1999] ausführlich beschreibt.

Seine Graphen sollen später zur Erkennung personenbezogener Prädikate in dem hier vorgestellten Ansatz eingesetzt werden. Da sie mit wenigen Ausnahmen alle Verben des Englischen in verschiedenen Zeitformen finden können, sind sie eine Bereicherung für jede Arbeit.

Die Abbildungen 3.2 und 3.3 auf Seite 26/27 visualisieren das Zusammenspiel der einzelnen Graphen. Dabei wird deutlich, welche Grammatik welche Grammatik aufruft, und es wird somit gezeigt, wie diese voneinander abhängen. Der Ausgangspunkt ist der Graph `VAUX`, der sozusagen alle Fäden bei der Erkennung der Verben „in der Hand hält“. Wie jeder dieser 80 Automaten genau aufgebaut ist, soll hierbei nicht von Interesse sein, da später (in Abschnitt 5.2.6) nur mit den von ihnen generierten Ergebnissen weitergearbeitet wird.

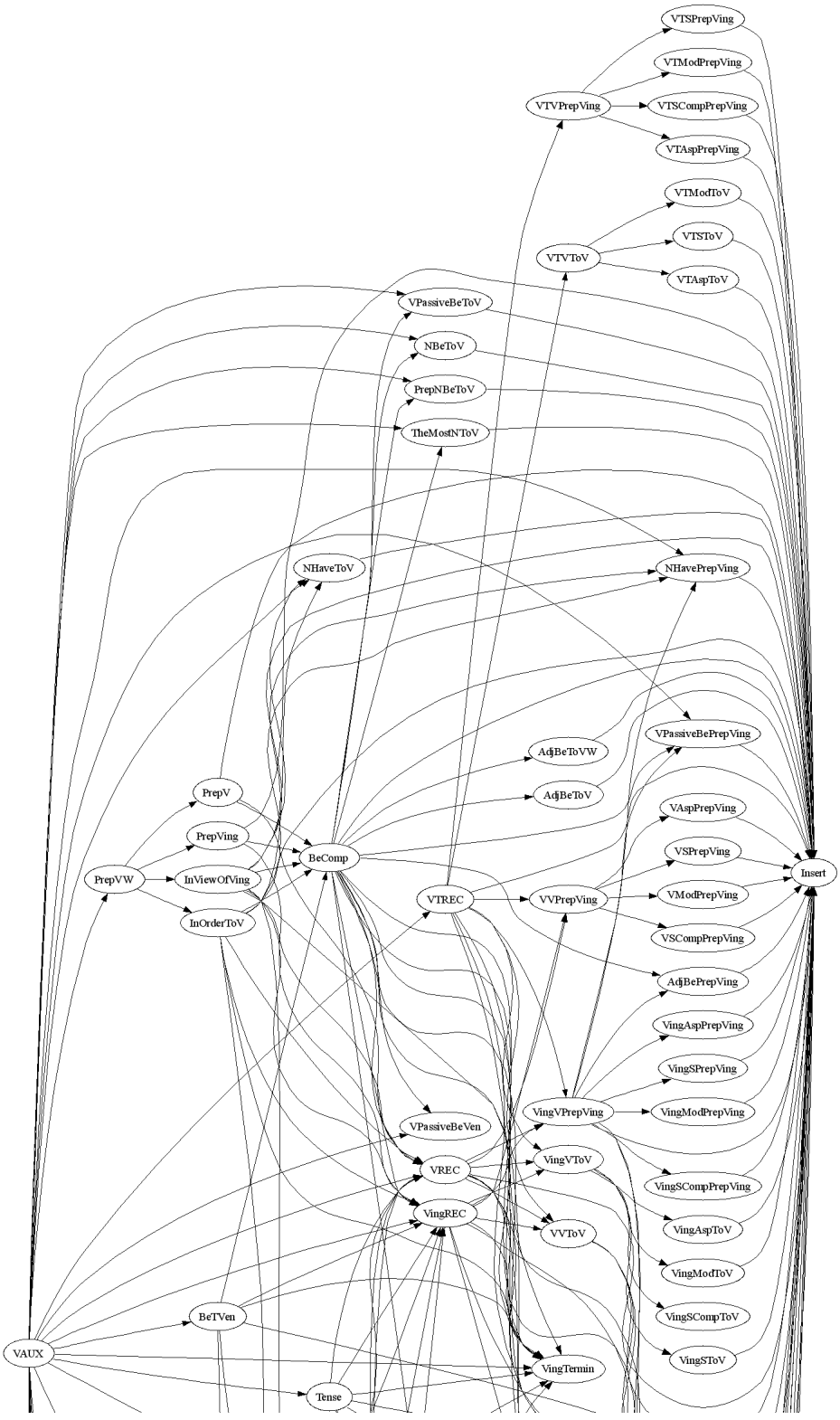


Abbildung 3.2: Übersicht der Lemmatisierungsgraphen aus [Gross, 1998-1999] (Teil 1)

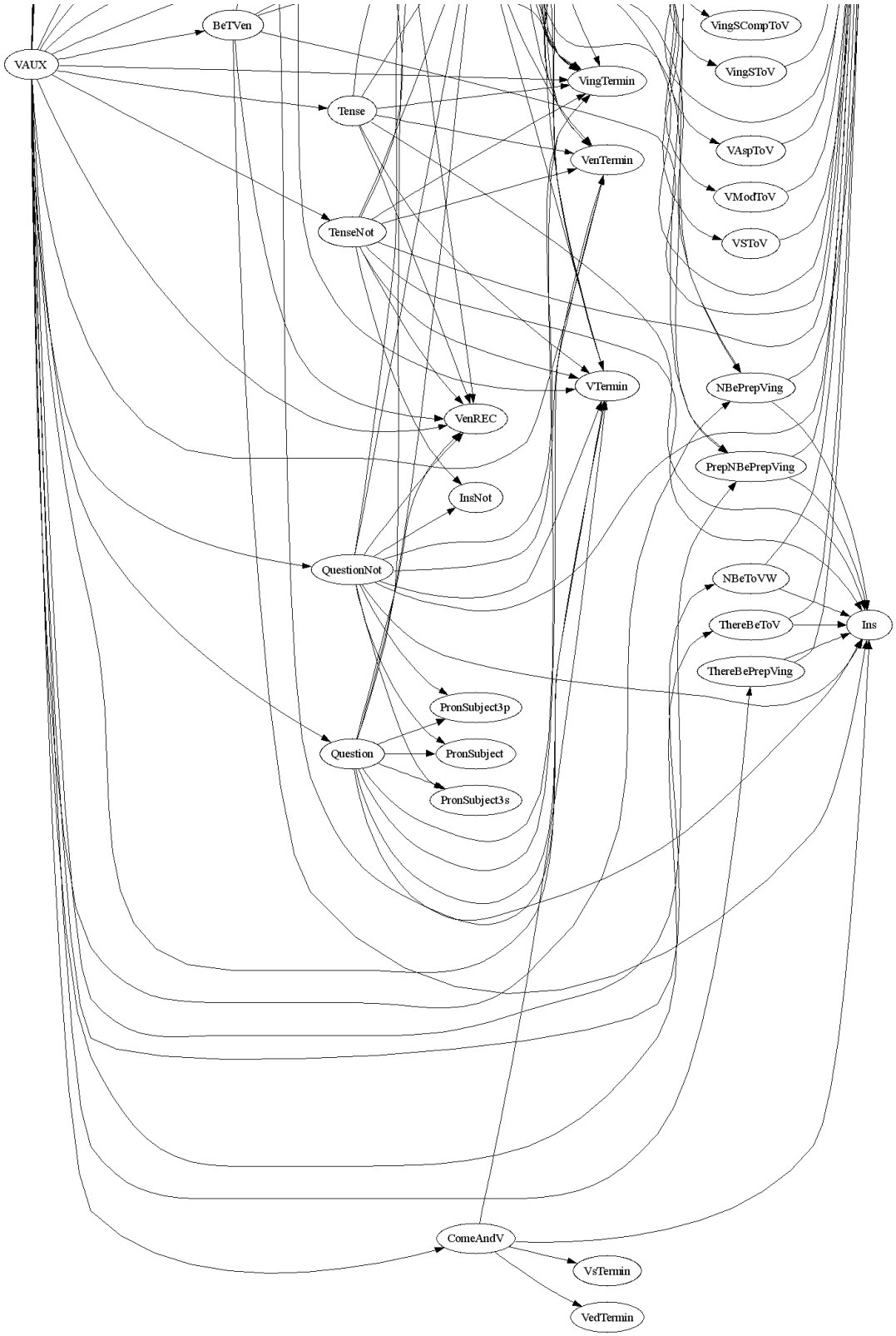


Abbildung 3.3: Übersicht der Lemmatisierungsgraphen aus [Gross, 1998-1999] (Teil 2)

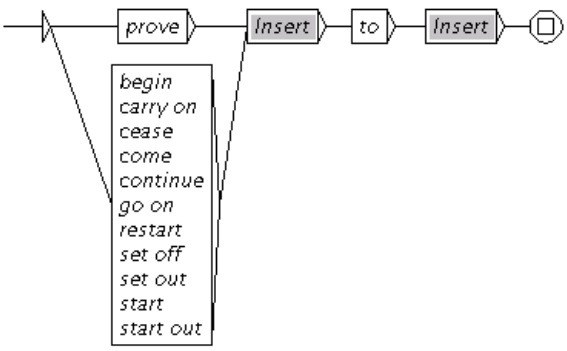


Abbildung 3.4: *VModToV.grf* aus [Gross, 1998-1999]

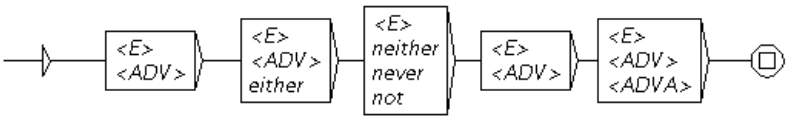


Abbildung 3.5: *Insert.grf* aus [Gross, 1998-1999]

Stellvertretend für alle Graphen veranschaulicht der Graph aus Abbildung 3.4 die Struktur einer möglichen Verbalphrase des Englischen und berücksichtigt dabei auch mögliche Satzeinschübe (siehe Abbildung 3.5).

3.3 Erkennung von Personenbezeichnungen

3.3.1 Erkennung von Eigennamen und Berufsbezeichnungen [Senellart, 1998a]

In seiner Arbeit *„Locating noun phrases with finite state transducers“* [Senellart, 1998a] beschreibt Jean Senellart einen wörterbuchgestützten Ansatz zur Erkennung von Eigennamen mittels endlichen Transduktoren.[11] Dafür hatte er sich zum Ziel gesetzt eine lokale Grammatik zu entwickeln, die Nominalphrasen bestehend aus Eigennamen und/oder Berufsbezeichnungen beschreibt. Jedoch sollte sich die Erkennung von Eigennamen – besonders von Personennamen – bzw. Berufsbezeichnern auf die Domäne der Zeitungsnachrichten beschränken.

Dabei müssen aber auch semantische Relationen, wie Synonymie und Hyperonymie berücksichtigt werden, so dass Anfragen vom Typ *„Find all newspaper articles in a general corpus mentioning the French prime minister.“* [Senellart, 1998a] oder *„How is Mr. X referred to in the corpus; what have been his different occupations through out the period over which our corpus extends?“* [Senellart, 1998a] verarbeitet werden konnten. Denn Antworten auf die erste Frage, werden wohl kaum Schlüsselworte aus der

[11] Endliche Transduktoren werden im Englischen als Finite-State-Transducers (FSTs) bezeichnet.

Query enthalten, sondern eher dazu passende Synonyme oder Eigennamen, die auf die Umschreibung „französischer Premierminister“ zutreffen.

Vergleich mit anderen Information Retrieval Methoden

Der eben beschriebene Ansatz weicht stark von anderen gängigen Ansätzen der Informationsbeschaffung aus unstrukturierten Texten ab. Weitere Konzepte zur automatischen Informationsgewinnung sind u.a.

- Algorithmen, die mit Schlüsselbegriffen arbeiten (Key-Word-Algorithms).
- Algorithmen, die nach Mustern exakt suchen (Exact-Pattern-Algorithms).
- Algorithmen, welche die Statistik zu Hilfe nehmen (Statistical Algorithms).

Key-Word-Algorithms werden gerne von Suchmaschinen, wie z.B. Yahoo!, verwendet. Sie suchen nach Schlüsselbegriffen aus der Anfrage, die zusammen in einem Text auftreten. In der Regel werden noch leichte Abwandlungen in der Rechtschreibung, sowie verschiedene grammatikalische Endungen und Rechtschreibfehler akzeptiert und bei der Suche miteinbezogen.
Exact-Pattern-Algorithms bzw. **Exact-String-Matching-Algorithms** verwenden reguläre Ausdrücke aus Buchstaben, welche exakt auf dem Dokument suchen. Mit dieser Methode arbeitet u.a. das Oxford English Dictionary (OED). Bei der Eingabe des Suchstrings sind jedoch auch Wildcards wie das Fragezeichen ? und der Asterisk $*$ erlaubt, wobei das Fragezeichen für einen beliebigen Buchstaben steht und der Asterisk eine beliebige Sequenz von Buchstaben repräsentiert. Des Weiteren beeinflusst die Groß- oder Kleinschreibung nicht das Auffinden von Einträgen, da *case-insentive* gesucht wird. Im Gegensatz zu den Key-Word-Algorithms muss jeder Term aus der Anfrage in der gegebenen Reihenfolge berücksichtigt werden.
Statistical Algorithms bieten dem Benutzer nur solche Dokumente als Ergebnis an, die sowohl Schlüsselwörter aus der Anfrage enthalten, aber auch statistisch gesehen semantisch nahe an den Anfragetermen liegen.

Am einfachsten zu implementieren sind wohl Algorithmen, die mit Schlüsselbegriffen aus der Anfrage arbeiten. Der Nachteil daran ist leider nur, dass die Ergebnisse sehr störanfällig sind, was nichts anderes heißt, als dass Homographen[12] der Anfrageterme im Text auftauchen können, oder dass Begriffe im Text gefunden werden, die sehr ähnlich zu den Anfragetermen sind.

Dagegen liefern Algorithmen, die mit Mustern bzw. regulären Ausdrücken arbeiten, ausgezeichnete Ergebnisse zurück. Jedoch sind die Muster hierbei so komplex, dass sich sogar Pattern spezifizieren lassen, mit denen man Synonyme der Anfrageterme finden kann. Außerdem lassen sich die verschiedenen grammatikalischen Endungen sehr präzise beschreiben. Nur wird es immer schwieriger die Muster zu konstruieren und zu verarbeiten, je komplexer die morphologischen Phänomene werden, welche es zu beschreiben gilt.

[12] Ein **Homograph** ist ein Wort, das die gleiche Schreibweise wie ein oder mehrere andere Wörter hat, aber von unterschiedlicher Bedeutung ist und meist auch unterschiedlich ausgesprochen wird.

Ein Algorithmus, der auf statistischen Methoden basiert, kann lediglich für einfache Anfragen gute Resultate liefern, und man braucht große Dokumentmengen, um statistisch repräsentative Ergebnisse zu bekommen. Doch dabei werden Terme mit niedriger Frequenz im Text meist ignoriert.

Welcher Ansatz wäre nun zur Erkennung von Eigennamen am idealsten?[13]

Der erste Ansatz würde funktionieren, wenn man entweder nur einen Vornamen oder einen Nachnamen in der Anfrage angeben würde, welcher auf keinen Fall ambig sein darf. D.h. dass beispielsweise Nachnamen wie *„Major“* nicht in der Query vorkommen dürfen, da sonst nicht nur ein Teil des Namens wie z.B. *„John Major“*, sondern auch *„Major Tom Stuart“* erkannt würde. Auch könnte dieser Algorithmus im Falle von mehreren Suchbegriffen, wie z.B. *John Major*, alle Artikel finden, in denen jemand erwähnt wird, der *„John“* heißt und alle Texte, in denen das Wort *„Major“* (als Eigenname oder als militärischer Rang) auftritt. Natürlich sollten auch Artikel gefunden werden, in denen beide Begriffe auftauchen, doch müssten sie nicht direkt nebeneinander im Text stehen, aber sie könnten es theoretisch. Die Implementierung des Algorithmus schreibt nicht vor, dass im Text zuerst *„John“* gefolgt von *„Major“* auftreten muss, was die Ergebnismenge für diese Zwecke unnötig vergrößert und die Präzision der Treffer deutlich verschlechtert.

Jedoch könnte womöglich der dritte Ansatz, welcher die Statistik miteinbezieht, relativ gute Antworten liefern, wenn man noch zusätzlich die Begriffe *prime* und *minister* mit in die Anfrage aufnehmen und auf sehr langen Dokumenten arbeiten würde. Dabei könnte man beispielsweise Nominalphrasen von der Art wie *„the prime minister, John Major“* oder *„the French prime minister“* extrahieren. Das sind äußerst zufriedenstellende Ergebnisse, wenn man an das anfänglich gesteckte Ziel – die Erkennung von Eigennamen – denkt. Somit ist der statistische Ansatz, der auf keinerlei grammatikalischen Beschreibungsmethoden basiert, nicht zu verachten.

Deshalb hat Jean Senellart zusammen mit Maurice Gross versucht, eine neue Methode zu entwickeln, die den statistischen Ansatz verbessert. In dem 1998 veröffentlichen Artikel *„Nouvelles bases pour une approche statistique.“* [Gross und Senellart, 1998] beschreiben sie die Möglichkeit einen Vorverarbeitungsschritt vor das statistische Matching zu schalten. Bei dieser Vorverarbeitung soll der Text zunächst nach Mehrwortlexemen – also mehreren Wörtern, die zusammen eine lexikalische Bedeutungseinheit bilden – durchsucht werden, so dass ungefähr 50% des Textes schon semantisch annotiert wurde. So kann es später beim statistischen Suchen auf dem Text nicht mehr möglich sein, dass z.B. die Wortgruppe *„prime minister“* oder *„energy minister“* bei der alleinigen Suche nach *„minister“* getrennt wird.

Obwohl diese erfolgreiche Zusammenarbeit von linguistischen mit statistischen Methoden einen sehr vielversprechenden Eindruck vermittelt, entschied sich Senellart bei seinem Vorhaben ganz auf die Dienste der Statistik zu verzichten und einen reinen grammatik- und wörterbuchgestützten Ansatz zur Erkennung von Eigennamen und Berufsbezeichnungen in Nominalphrasen zu wählen.
Auf der Basis großer Lexika mit Eigennamen und Berufsbezeichnungen und unter Verwendung von Transduktoren sollten Grammatiken für die englische Sprache entstehen,

[13] vgl. [Senellart, 1998a]

welche Satzteile mit Personennamen oder Berufsbezeichnern formal und vollständig beschreiben.

Funktionsweise des Algorithmus

Der Algorithmus lässt sich in drei große Verarbeitungsschritte unterteilen.

1. Zunächst werden die Wörterbücher für die Eigennamen, sowie die lokalen Grammatiken, welche die Berufsbezeichungen beschreiben, auf das Korpus angewendet. Dabei werden semantische Relationen wie Synonymie und Hyponymie und die Zeitlinie der Textsammlung formal definiert. Damit man die Ergebnisse dieses Schrittes in Echtzeit zurückgeliefert bekommt, wird auf einem zuvor konstruierten Index der Datensammlung gearbeitet.

2. In dieser Phase werden die erkannten Eigennamen im Transduktor durch Variablen ersetzt und die gefundenen Eigennamen werden zur Lokalisierung anderer Eigennamen verwendet, die dann dem Benutzer als neue Wörterbucheinträge angeboten werden. Dadurch kann das Erstellen von weiteren Transduktoren und die Ergänzung der Lexikoneinträge überwiegend automatisiert werden.

3. Zum Schluss werden die erkannten Nominalphrasen automatisch in andere (natürliche) Sprachen übersetzt, indem entsprechende Transduktoren für die jeweilige Sprache generiert werden.

Einblick in die formalen Beschreibungsmethoden

Abbildung 3.6 zeigt eine lokale Grammatik in Form eines Finite-State-Graphen (FSG)[14]. Ein FSG ist im Grunde nur die graphische Repräsentation eines Finite-State-Transducers (FST). Jeder einzelne Knoten stellt die jeweilige Eingabesequenz dar, die der Automat an dieser Transition akzeptiert. Unterhalb mancher Knoten befinden sich Markierungen, welche die Ausgabesequenzen für den entsprechenden Input im Knoten darüber illustrieren. Der Startzustand des Transduktors wird durch einen Linkspfeil markiert, wohingegen der Endzustand als doppeltes Quadrat angedeutet wird. Hat ein Knoten einen leicht grauen Hintergrund, so heißt das, dass er einen Subtransduktor aufruft – eine Schreibweise, die es ermöglicht, die Übersichtlichkeit der Automaten zu gewährleisten. Natürlich ist es auch möglich, dass ein Subgraph einen Output hat, der dann in die Ausgabe des Haupttransduktors miteinbezogen wird. Ein Knoten, der ein `<E>` beinhaltet, symbolisiert die leere Transition.

Mithilfe dieser Darstellungsformalismen lassen sich linguistische Konstrukte recht einfach darstellen, da z.B. das System UNITEX auch einen Graphen-Editor bietet, mit dem sich solche Grammatiken leicht erstellen lassen. Außerdem sind diese FSTs besser als gewöhnliche FSTs, da die Subgraphen sich auf den Hauptgraphen – also auf den Kontext davor oder danach – beziehen können, so dass man mit ihnen auch kontextsensitive Wörter[15] des Typs $a^n b^n$ erkennen kann.

[14] vgl. [Senellart, 1998a]
[15] siehe [Hopcroft *et al.*, 2002]

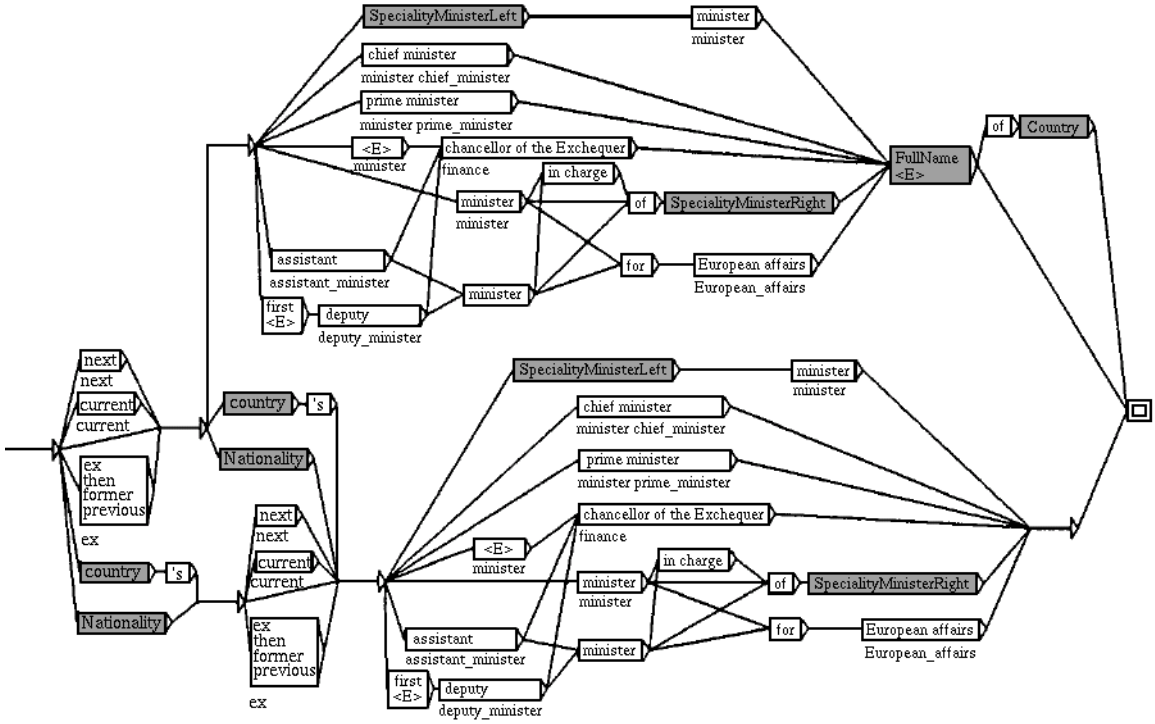


Abbildung 3.6: *MinisterOccupation.grf* aus [Senellart, 1998a]

An diesem konkreten Beispiel aus Abbildung 3.6 soll die Problematik der formalen Beschreibung von Nominalphrasen, welche sich auf das Wort *„minister“* beziehen, behandelt werden. Dieser Graph erkennt beispielsweise die Sequenz *„minister for European affairs“*, aber er würde nicht *„French minister for agriculture“* matchen. Somit wäre dieser Graph sicher noch ausbaufähig.

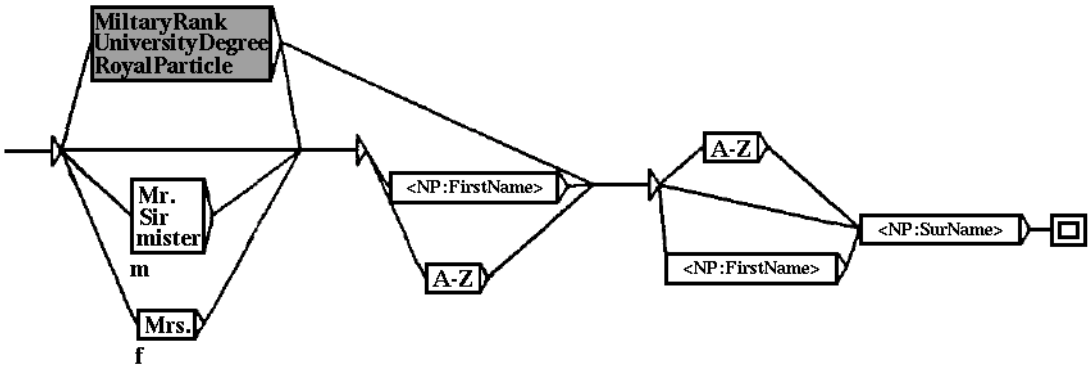


Abbildung 3.7: *FullName.grf* aus [Senellart, 1998a]

Der Graph aus Abbildung 3.7 illustriert die Erkennung von Personennamen, wobei er sich „Wörterbuch-Look-Ups“ zunutze macht.

An Knoten, welche `<PN:FirstName>` oder `<PN:SurName>` enthalten, gleicht der Transduktor alle potentiellen Vornamen oder Nachnamen aus den Lexika mit dem Text ab. Deshalb ist die Ausgabe dieses Automaten ein Nachname, vielleicht noch ein Vorname, und wenn vorhanden *„m“* oder *„f“* für das ermittelte Geschlecht der Person, wobei das Geschlecht über die Anrede *„Mr*, *Sir*, *mister*, *Mrs“* ermittelt wird.

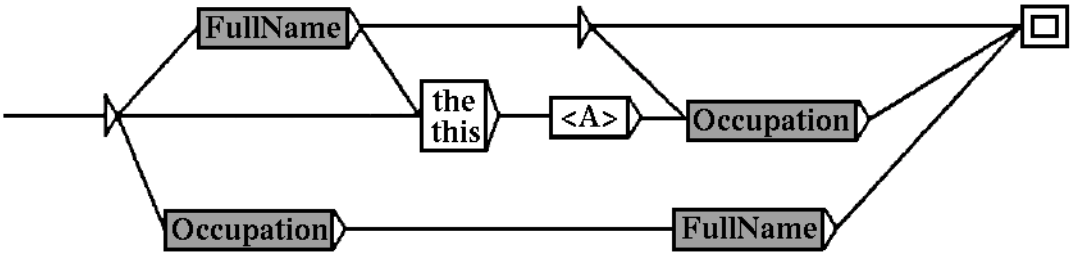


Abbildung 3.8: *NounPhrases.grf* aus [Senellart, 1998a]

Der `NounPhrases`-Graph in Abbildung 3.8 vereinigt die Subgraphen *Occupation.grf* und *FullName.grf* und stellt somit die syntaktische Beziehung zwischen diesen beiden semantischen Klassen der Eigennamen und Berufsbezeichnungen her. Dabei ist anzumerken, dass `<A>` stellvertretend für alle Adjektive steht, die dem Standardwörterbuch bekannt sind. Somit würde dieser Automat u.a. Phrasen wie *„software engineer Tom Mitchell“* oder *„Harry Smith the fantastic cook“* erkennen.

Schwächen des Ansatzes

Natürlich beschreibt das komplette Graphenpaket nicht alle syntaktischen Möglichkeiten, wie Personennamen zusammen mit Berufsbezeichnungen auftreten können. Dennoch versucht es nahezu alle einfachen Konstruktionen abzudecken. Beispielsweise würde der `NounPhrases`-Graph aus Abbildung 3.8 nicht auf dem Satz *„Mr. Smith, who is since 1978, the chairman of ...“* matchen, da der Nebensatz im Graphen nicht berücksichtigt wird. Auch andere Einschübe dieser Art sind kompliziert zu erfassen und werden in diesem Automaten außer Acht gelassen.

Eine andere Schwierigkeit besteht darin, dass eine Person mehrere Berufe ausüben kann, und somit besteht keine Möglichkeit, eine eindeutige Zuordnung zwischen Person und Beruf zu machen. Denn sie wird eventuell an einer Stelle im Text mit einer Berufsbezeichnung und an einer anderen Position im Dokument mit einem anderen Beruf referenziert. Dadurch kann das System nicht gewährleisten, dass bei folgender Zuordnung

```
SurName=Mitchell, FirstName=Tom, Gender=m, Occupation=cook
SurName=Mitchell, FirstName=Tom, Gender=m, Occupation=hotel manager
```

„Tom Mitchell“ ein und dieselbe Person ist.

Auch sind Adverbiale, die an fast jeder Stelle im Satz auftreten können, schwer einer Berufsbezeichnung als deren Ergänzung zuzuordnen. So ist es selbst im Beispiel *„In China, the first minister has ...“* für den menschlichen Betrachter schwierig, die Ortsergänzung *„In China“* der Berufsbezeichung *„first minister“* zuzuordnen.

3.3.2 Erkennung von Personennamen in Zeitungstexten [Friburger, 2002]

Nathalie Friburger hat sich im Zuge ihrer Dissertation der Erkennung von Eigennamen in Nachrichtentexten gewidmet. Ähnlich wie Jean Senellart [Senellart, 1998a; Senellart, 1998b] wählt sie einen wörterbuchgestützten Ansatz zur Erkennung von Eigennamen mittels Transduktoren. Um einen kurzen Überblick zu geben, wie sie an dieses Thema herangeht, soll nun die Arbeit von ihr und Denis Maurel *„Elaboration d'une cascade de transducteurs pour l'extraction des noms personnes dans les textes“*[16] [Friburger und Maurel, 2001] und ihre Doktorarbeit *„Reconnaissance automatique des nomes propres – Application à la classification automatique de textes journalistiques“*[17] [Friburger, 2002] vorgestellt werden.

Kaskadierung[18] von Transduktoren zur Extraktion von Eigennamen [Friburger und Maurel, 2001]

Hierbei handelt es sich um ein Programm, welches Personennamen in französischen Zeitungsberichten erkennt.

Transduktoren sind im Grunde auch nur endliche Automaten, welche allerdings ein Eingabe- und ein Ausgabealphabet haben. In diesem Fall besteht das Eingabealphabet aus Mustern, die im Korpus gefunden wurden, und das Ausgabealphabet fügt den mit den Pattern erkannten Passagen die passende XML-Information hinzu. In der Regel sind die gefundenen und XML annotierten Sequenzen Personennamen und ihre jeweiligen Kontexte, was folgendes Beispiel illustriert.

Le Juge Renaud Van Ruymbeke

⇓

<profession> *judge* <\profession>

<person> <prenom> *Renaud* <\prenom> <nom> *Van Ruymbeke* <\nom><\person>[19]

Doch bevor man die Transduktoren nacheinander auf das Korpus anwenden kann, sind einige Vorverarbeitungschritte notwendig, welche vom System INTEX [Silberztein, 1993] übernommen werden.[20] Zu den wichtigsten Phasen zählt u.a. die Satzenderkennung, welche auf dem ganzen Text durchgeführt wird. Im gleichen Schritt werden auch die Satzendmarkierungen in den Originaltext eingefügt. Später erfolgt die Anwendung sämtlicher Wörterbücher auf das Korpus, wobei jedes Wort mit allen Formen, die in einem der Lexika auftreten, markiert wird. An dieser Stelle wird noch *keine* Disambiguierung durchgeführt. Jedoch besitzen die annotierten Wörter nun sämtliche grammatikalischen und semantischen Informationen, die in den Lexika kodiert sind.

[16] englische Übersetzung des Titels: *„Finite-state transducer cascades to extract named entities in texts“*.

[17] englische Übersetzung des Titels: *„Automatic Recognition of Proper Names – An Application in Automatic Clustering of Journalistic Texts“*.

[18] In diesem Fall bedeutet Kaskadierung das Zusammenschalten von verschiedenen Transduktoren (Reihenschaltung von Transduktoren).

[19] aus [Friburger und Maurel, 2001]

[20] Das System UNITEX [Paumier, 2004] bietet ähnliche Funktionen zur Korpusbearbeitung.

Dabei kamen die Wörterbücher

- *DELAS* – ein Lexikon, welches die gesamte grammatikalische Information für „einfache Wörter“[21] festhält,
- *Prolintex*, ein Toponymlexikon[22],
- *Prenom-prolex*, ein Wörterbuch für Vornamen,
- und ein Lexikon für Berufsbezeichnungen, das von Cédrick Fairon an der Universität Marne-la-Vallée erstellt wurde.

zum Einsatz.

Das Prinzip der Kaskadierung von Transduktoren ist im Grunde recht einfach zu erklären. Die Transduktoren müssen in einer aufeinander abgestimmten Reihenfolge nacheinander auf den Text angewendet werden. Denn oft ist der Output eines Transduktors, der Input – das zu Suchende – für den darauffolgenden Transduktor. Jede gefundene Sequenz wird markiert (siehe Beispiel Seite 34) und kann durch diese Markierung über den Index gefunden werden. Auch muss jedes erkannte Muster aus dem Text gelöscht werden, da sonst die Gefahr besteht, dass ein später geschalteter Transduktor es nochmal erkennt. So wird vermieden, dass Passagen mehrfach erkannt werden, und dass das System ineffizient arbeitet.

Eine Voraussetzung muss noch erfüllt werden, bevor die Transduktoren in Reihe geschaltet werden können. Es ist auch wichtig, sich eine Sammlung an linken und rechten Kontexten der Personennamen, die in dem Zeitungskorpus vorkommen, aufzubauen. Denn das Matchen über die Kontexte von Personennamen stellte sich bei französischen Texten als äußerst hilfreich heraus, weil ungefähr 90% aller Personennamen in Nachrichtentexten über ihren linken Kontext erkannt werden können. Ein Grund dafür könnte sein, dass gewisse Stilkonventionen zur Behandlung von Personennamen in Printmedien bestehen, so dass eindeutige, fast standardisierte Muster erkennbar waren.

Mithilfe eines annotierten Korpus der französischen Zeitung *Le Monde*, der ungefähr 165000 Wörter umfasste (*Ouest France* enthielt 67000 Wörter) war es möglich, die häufigsten Kontexte von Personennamen im Text zu kategorisieren.

- In 25,9% (17,1% für *Ouest France*) der Fälle ging dem Personennamen ein Titel oder eine Berufsbezeichnung gefolgt von einem Vornamen und/oder einer Staatsangehörigkeit voran. **(Fall 1)**
- In 19,1% (16,3% für *Ouest France*) der Fälle ging dem Nachnamen ein Berufsbezeichner oder ein Titel zusammen mit einer Bezeichnung für eine Staatsangehörigkeit oder ein dem Wörterbuch unbekannter Vorname zusammen mit einer Nationalität voran. **(Fall 2)**

[21] „simple words“ - im Gegensatz zu „compound words“ (Mehrwortlexemen), welche das *DELAC* auflistet.

[22] Ein **Toponym** bezeichnet einen Ortsnamen im allgemeinen Sinne. Hierunter versteht man also insbesondere die Bezeichnungen bestimmter Gebiete, Verwaltungseinheiten, Siedlungen, Verkehrswege, Gewässer und alle übrigen topographischen Objekte mit Eigennamen. [vgl. `http://de.wikipedia.org/wiki/Toponym`]

- Am häufigsten mit 43,4% (59,0% für *Ouest France*) hat der Kontext eher eine beschreibende Funktion zum jeweiligen Personennamen, d.h. er wird meist attributiv eingesetzt und der Personenname besteht in der Regel aus einem dem Lexikon bekannten Vor- und Nachnamen. **(Fall 3)**

- Setzt man den Kontext zur Erkennung der Eigennamen ein, so helfen Berufsbezeichnungen oder Verben der Äußerung, wie z.B. *„sagen“* oder *„erklären“* dabei, 5,2% (2,2% für *Ouest France*) aller Personen im Text zu erkennen. Natürlich können Verben der Äußerung auch ohne ein menschliches Subjekt im Satz auftreten, d.h. diese Kontexte sind mit Vorsicht zu genießen. **(Fall 4)**

- Die übrigen 6,4% (5,4% für *Ouest France*) der Personennamen weisen keinerlei hilfreiche Kontexte auf, so dass diese nutzlos bei der Suche sind. Diese Personen sind in der Regel so berühmt, dass der jeweilige Autor es wahrscheinlich nicht für nötig gehalten hat, die Persönlichkeit vorzustellen, oder ein paar einleitende Worte zu ihr zu schreiben. Doch ca. die Hälfte dieser anscheinend nicht im Text zu findenden Leute, werden an anderen Stellen im Korpus nochmal namentlich erwähnt, so dass im Endeffekt nur noch 3,3% der ursprünglich 6,4% Personennamen unerkannt bleiben. Eventuell könnte ein Lexikon aller berühmten Personennamen diesen Prozentsatz weiter verringern. **(Fall 5)**

Dass die Trefferquoten im ersten und im zweiten Fall für *Ouest France* kleiner als für *Le Monde* ausfallen, liegt wohl an strikteren Schreibkonventionen, die für die Journalisten von *Le Monde* bestehen. Somit können die vordefinierten Muster erfolgreicher auf *Le Monde* als auf *Ouest France* suchen.

Um die gute Trefferquote ihres Ansatzes der Kaskadierung von Transduktoren nachzuweisen, wandte Nathalie Friburger 14 Transduktoren in Reihe geschaltet nacheinander auf ein Teilkorpus von *Le Monde* an, das etwa 80.000 Wörter umfasste.

Dabei ergaben sich je nach Fall (siehe oben) folgende Ergebnisse:

	Fall 1	**Fall 2**	**Fall 3**	**Fall 4**	**Fall 5**	**Gesamt**
Tatsächliche Anzahl der Personennamen im Text	253	187	424	50	64	977
Anzahl der gefundenen Personennamen im Text	245	187	413	32	32	909
Anzahl der korrekt gefundenen Personennamen im Text	242	186	410	30	31	899
Recall	95,7%	99,5%	96,7%	60,0%	48,4%	91,9%
Precision	98,8%	99,5%	99,3%	93,8%	96,9%	98,7%

Mit den Resultaten in den ersten der drei Fälle kann man sehr zufrieden sein. Doch leider weist Fall 4 einen schlechten Recall auf, was wohl mit der problematischen Erkennung von Eigennamen in ambigen Kontexten zu tun hat. Fall 5 behandelt nur Namen, die ohne einen spezifischen Kontext im Korpus auftreten, und die nur durch die Gesamtbedeutung des Satzes oder durch das Wissen eines menschlichen Lesers identifiziert werden können.

Da einige dieser Personen schon in anderen Textpassagen vorkamen, ist es überhaupt möglich einen Recall von 48,4% zu erreichen.

Abschließend kann man sagen, dass das Prinzip der Kaskadierung von Transduktoren recht einfach und effektiv bei der Suche nach Personennamen sein kann. Dagegen gehen Nathalie Friburger und Denis Maurel davon aus, dass die Extraktion von anderen Eigennamen, wie Orts- und Organisationsnamen wesentlich schwieriger ist, weil ihre jeweiligen Kontexte im Korpus nicht so schematisch wie die von Personennamen sind.

Eigennamen bei der Klassifikation von Nachrichtentexten

Nathalie Friburger hat die Idee der Reihenschaltung von FSTs auch für ihre Dissertation eingesetzt. Vollständigkeitshalber soll noch kurz die Thematik ihrer Doktorarbeit angesprochen werden, bei der das System *casSys* zum Einsatz kam, welches die Kaskadierung von Transduktoren implementiert. Das Ziel ihrer Arbeit war nicht nur die automatische Extraktion von Eigennamen, sondern auch die automatische Klassifikation von Zeitungstexten anhand der darin auftretenden Namen. Das dafür eingesetzte Programm *extractNP*, welches *casSys* verwendet, ermöglicht es Ambiguitäten aufzulösen, sowie Eigennamen zu segmentieren und kategorisieren. Das System lieferte hervorragende Ergebnisse, so dass eine Präzision von 94% und ein Recall von 93% erzielt wurde. Des Weiteren entwickelte sie eine Anwendung, welche sich die verschiedenen Vorkommen von Personennamen zunutze macht, um Zeitungsnachrichten nach Thematiken zu kategorisieren. Dabei stellte sich heraus, dass dieser Ansatz ein qualitativ gutes Clustering von Zeitungstexten ermöglichte.

3.4 Erkennung von Organisationsnamen in Wirtschaftsnachrichten [Mallchok, 2004]

In ihrer Doktorarbeit *„Automatic Recognition of Organization Names in English Business News“*[23] hatte sich Friederike Mallchok zum Ziel gesetzt, nachzuweisen, dass sich die Genauigkeit und Performanz der Eigennamenerkennung wesentlich verbessern lässt, wenn man einen sprachspezifischen Ansatz dafür wählt. Unter einem sprachspezifischen Ansatz versteht man einerseits die Beschränkung der Trainingskorpora auf eine bestimmte Domäne, wie z.B. den Bereich der Wirtschaftsnachrichten, und andererseits aber auch eine Einschränkung bei der Named Entity Recognition (NER). Wenn man eine *Named Entity* (benannte Entität), wie hier die Organisationsnamen, in den Vordergrund rückt, und dann ausgehend von dieser bestimmten Klasse der Eigennamen ihre Kontexte untersucht, finden sich weitere Eigennamen und noch weitere wertvolle Informationen in ihrem Umfeld. Um die Kontexte der Organisationsnamen syntaktisch und semantisch beschreiben zu können, wählte Friederike Mallchok die formale Repräsentation der lokalen Grammatiken. Dabei verzichtet sie vollständig auf statistische Methoden zur Extraktion von Eigennamen und verlässt sich ganz auf die Identifikation von Organisationsnamen durch ihre jeweiligen Kontexte und das in Wörterbüchern kodierte Zusatzwissen.

[23] Automatische Erkennung von Organisationsnamen in Englischsprachigen Wirtschaftsnachrichten

Einsatz von Ressourcen: Korpora und Lexika

Da Friederike Mallchok sich dazu entschlossen hatte, die für Wirtschaftsnachrichten typische „Subsprache“[24] zu untersuchen, fiel ihre erste Wahl auf das frei verfügbare Reuters Korpus[25]. Das Reuters Korpus enthält alle Nachrichtentexte (ca. 810.000), welche die Nachrichtenagentur Reuters Ltd. vom 20. August 1996 bis einschließlich 19. August 1997 veröffentlicht hatte.

Nachdem dieses Korpus für Wirtschaftsnachrichten nicht mehr auf dem neuesten Stand war, ergänzte sie ihre Textsammlung durch Online-Ausgaben der *„Financial Times“*, des *„Wall Street Jounals“*, von *„Newsday“*, der *„New York Times“* und durch aktuelle Artikel der *„Reuters News“*. Denn gerade für die Erkennung von Firmennamen ist es wichtig, aktuelle Informationen über die Unternehmen vorliegen zu haben. Beispielsweise können einerseits dem Lexikon bekannte Firmen, welche nach 1997 gegründet wurden, nicht im Reuters Korpus gefunden werden, und andererseits werden nur Namen von Organisationen in diesem Text lokalisiert, welche in dieser Zeitspanne in den Nachrichten präsent waren. Unter Hinzunahme der eben angesprochenen elektronischen Nachrichtenausgaben konnten auch junge, aufstrebende oder immer noch bedeutende Unternehmen in den aktuellen Texten erkannt werden.

Wie eben kurz erwähnt, wurden mehrere semantische Lexika unterstützend zur Erkennung der Organisationsnamen in den Korpora eingesetzt. Mittels dieser Wörterbücher sollte das Auffinden von Firmennamen im Text wesentlich erleichtert werden.

Mithilfe diverser Internetressourcen konnte Friederike Mallchok ein beachtliches Begriffsinventar für ihr Organisationsnamenlexikon (ONL) und für ihr Organisationsbeschreibungslexikon (ODL) zusammenstellen. Die jeweiligen Namen der Lexika lassen natürlich schon auf ihren Inhalt schließen: Das ONL enthält ausschließlich Firmennamen, und das ODL führt eine Reihe an Beschreibungen für Unternehmen auf, welche oft in Wirtschaftstexten den Organisationsnamen einleiten.

Zudem ließen sich im Kontext von Organisationsnamen relativ oft Berufsbezeichner finden, welche in einem Berufsbezeichnerlexikon (HPL) archiviert wurden. Des Weiteren konnten Ortsbezeichnungen wie Länder, Städte und Staaten, sowie Zeitangaben in den entsprechenden Lexika gespeichert werden. Auch allgemeine Kontexte der Firmennamen extrahierte Friederike Mallchok aus den Korpora und bewahrte sie in Wörterbüchern auf. Dabei wurden nur die textuellen Umgebungen von Organisationsnamen ins Lexikon übernommen, welche besonders häufig in den Korpora vorkamen.

Entwicklung lokaler Grammatiken

Basierend auf den eben genannten Lexika entwickelte sie lokale Grammatiken, welche einerseits die interne Struktur von Organisationsnamen repräsentierten und andererseits auf ihre Funktion im Satz eingehen bzw. ihr syntaktisches Verhalten in Wirtschaftsnachrichten wiederspiegeln. Die verschiedenen Grammatiken sollten so viele syntaktische Variationen wie möglich abdecken, in denen Unternehmen vorkommen können. Somit

[24] Einschränkung der Sprache auf eine bestimmte Bezugsdomäne wie z.B. Wirtschaftsnachrichten, sowie Dominanz von Fachvokabular.

[25] `http://about.reuters.com/researchandstandards/corpus/`

werden in diesen lokalen Grammatiken Möglichkeiten berücksichtigt, Organisationsbeschreibungen den Firmennamen voran- oder nachzustellen, sowie Berufsbezeichner – eventuell in Verbindung mit einem Personennamen – im linken oder rechten Kontext der Organisationsnamen zu nennen. Den Schwerpunkt ihrer Studien legte sie auf die Wirtschaftereignisse *„joint venture“*, *„merger“* und *„partnership“*, für die sie zeigen wollte, dass eine erweiterte Indexierung durch lokale Grammatiken, welche diese Phänomene beschreiben, durchaus möglich ist und später für eine intelligente und effektive Suche eingesetzt werden kann.[26]

Bootstrapping und Akronymbildung

Wie bereits Maurice Gross (siehe Abschnitt 3.1.1) und Jean Senellart (siehe Abschnitt 3.1.2) sich des Bootstrappings bei der Entwicklung von Transduktoren (Finite-State-Transducern) bzw. lokaler Grammatiken mit dem System UNITEX [Paumier, 2004] bedient haben, verwendet auch Friederike Mallchok diese Methode zur Verbesserung ihrer Ergebnisse. Auf diese Weise stellte sich in mehreren Nachbearbeitungsschritten schnell heraus, welche Fehlerquellen noch in den Grammatiken vorlagen, und wie diese minimiert werden konnten. Zusätzlich generierte sie aus den Organisationnamen, die aus mehreren Wörtern zusammengesetzt waren, mögliche Akronymvarianten, welche später im Korpus verifiziert wurden. Außerdem wurden noch weitere Abkürzungsmöglichkeiten für die entsprechenden Firmennamen berücksichtigt und ihre Existenz auf dem Text überprüft. Bei der erfolgreichen Validierung wurden die Varianten der Organisationsnamen in das Lexikon aufgenommen und im Korpus annotiert.

Fazit der Arbeit

Mit dem entwickelten System ist es Friederike Mallchok gelungen, Organisationsnamen mit einer hohen Genauigkeit und guten Performanz in englischsprachigen Wirtschaftsnachrichten zu erkennen. Dabei war es ihr möglich zu zeigen, dass das Ergebnis der Eigennamenerkennung signifikant verbessert werden kann, wenn jede Sprache, jede Domäne und jede Art von Entität getrennt behandelt wird. Außerdem widerlegte sie die Annahme von vergleichbaren NER[27]-Systemen, dass die Verwendung von Kontextinformationen nur zur Lokalisierung von Entitäten sinnvoll ist. Ihr Ansatz bewies, dass durch den Einsatz von lokalen Grammatiken weitere Informationen über die entsprechenden Entitäten aus den Korpora gewonnen werden können. So dienen beispielsweise semantisch kategorisierte Organisationsbeschreibungen dazu, die Entitäten, die sie beschreiben oder sogar die Texte oder Textabschnitte, in denen diese vorkommen, zu klassifizieren.
Ihre Bemühungen auf dem Gebiet der automatischen Erkennung von Organisationsnamen in Wirtschaftsnachrichten brachten sie letztendlich zu dem Schluss, dass eine Weiterentwicklung dieser lokalen Grammatiken auf jeden Fall sinnvoll ist. Dadurch könnte später eine breitere Abdeckung auf der Domäne der Wirtschaftnachrichten erreicht werden.

[26] vgl. `http://www.cis.uni-muenchen.de/~schmidt/lg/Deutsche_Zusammenfassung.pdf`
[27] Named Entity Recognition

4 Beschränkungen im System

Wie bereits Friederike Mallchok so treffend in ihrer Doktorarbeit [Mallchok, 2004] bemerkt hatte, ist einer der größten Vorteile von lokalen Grammatiken die Modularität. Es ist keinenfalls ein Nachteil sich bei der Erstellung lokaler Grammatiken besonders auf eine bestimmte Entität zu konzentrieren und deren Kontext möglichst genau zu beschreiben.

Der hier vorgestellte Ansatz fokussiert zwar die Erkennung von Menschenbezeichnern und beschränkt sich auf biographische Relationen (siehe Abschnitt 1.2), doch wird sich bei der Entstehung des Systems zeigen, dass auch andere Entitäten in der Umgebung von Personen auftreten und somit berücksichtigt werden müssen. Dafür wurden lokale Grammatiken entwickelt, die Personen näher spezifizieren, aber auch Organisationen und Toponymen eine gewisse Beachtung schenken. Überdies werden auch Verbrelationen in Form von lokalen Grammatiken beschrieben, welche die verschiedenen Entitäten miteinander logisch und syntaktisch verbinden.

So kann jede Grammatik separat erweitert und auch die darin verwendeten semantischen Hyperonymklassen können jederzeit durch weitere Wörterbucheinträge ergänzt werden. Des Weiteren lassen sich die entstandenen Grammatiken problemlos in andere NER-Systeme integrieren. Auch zur Erkennung anderer Entitäten sollten die Informationen aus diesen lokalen Grammatiken herangezogen werden, so dass man auf diesem Wissen aufbauen und zugleich das System erweitern könnte.

Außerdem sollte die Entscheidung, sich bei der Erkennung von Personen innerhalb biographischer Kontexte auf die Domäne der Wirtschaftsnachrichten zu beschränken, kein Hindernis dafür sein, später die für diesen speziellen Bereich entwickelten Grammatiken für andere Themengebiete auszuweiten. Denn wie die Wahl meiner Korpora zeigen wird, gibt es biographische Relationen, welche äußerst selten in Wirtschaftsnachrichten auftreten, dagegen aber in einer richtigen Biographie kaum fehlen. Es ist nur natürlich, dass nicht jedes personenbezogene Prädikat in einem Wirtschaftstext eine biographische Relation verkörpert, und dass nicht jede biographische Relation in den Nachrichten veröffentlicht wird. Das ist für die Entwicklung lokaler Grammatiken nur insofern ein Problem, wenn Verbrelationen beschrieben werden, welche höchst selten im Trainingskorpus vorkommen. Somit ist die Qualität einer lokalen Grammatik schwer zu messen und alternative Trainingskorpora werden benötigt. Man sollte sich dieser Tatsache immer bewusst sein, dass die Entwicklung lokaler Grammatiken stark von der Domäne des Korpus abhängt und sein Einfluss auf die Grammatik nicht zu unterschätzen ist. Auch wenn der hier präsentierte Ansatz sich hauptsächlich auf biographische Relationen konzentriert, die häufig in Wirtschaftsnachrichten vorkommen, soll das nicht heißen, dass diese lokalen Grammatiken nicht auf Texten anderer Bereiche gute Ergebnisse erzielen. Es werden lediglich die Relationen nicht abgedeckt, die kaum oder nie in Wirtschaftstexten genannt werden, was ein Ansporn wäre, das Konzept auszuweiten.

4.1 Sprachgebundenheit

Alle hier vorgestellten lokalen Grammatiken wurden für die englische Sprache entwickelt.

Sicherlich ist die Entscheidung, für welche Sprache die Erkennung von Personen in biographischen Kontexten implementiert wird, nicht unbegründet getroffen worden. So wurde die Wahl der Sprache sicher durch die große Dominanz des Englischen als Sprache des Internets beeinflusst. Doch auch die Tatsache, dass für das Englische schon sehr viel im Bereich Named Entity Recognition (NER) erforscht und entwickelt worden ist, wovon man manches aufgreifen, verbessern oder mit seinem eigenen Ansatz vergleichen kann, spielte eine beachtliche Rolle bei dieser Entscheidung.

Soweit es das Gebiet der lokalen Grammatiken betrifft, wurde die meiste Vorarbeit bei der linguistischen Analyse der französischen und englischen Sprache geleistet.

Außerdem ist der Bereich der Wirtschaft ein von Anglizismen geprägtes Feld, was ebenfalls dafür sprechen würde, sich gleich auf die Originalsprache zu konzentrieren.

Überdies ist die Auswahl an Trainingskorpora wesentlich größer, wenn man sich für die Arbeit mit Englisch entscheidet, und bei der Erstellung von Lexika kann im Internet auf ein großes Spektrum an Ressourcen in Form von themenspezifischen Listen zurückgegriffen werden, so dass für das Englische in kürzerer Zeit als für eine andere Sprache eine enorme Wissensbasis zusammengestellt werden kann.

Trotz der Beschränkung auf das Englische bei der Entwicklung lokaler Grammatiken, können die entstandenen Grammatiken mit relativ wenig Aufwand auf andere Sprachen übertragen werden.

4.2 Schwerpunkt Wirtschaftsnachrichten

Für computerlinguistische Untersuchungen wurden immer schon gern Korpora herangezogen, welche aus Wirtschaftstexten zusammengestellt waren. Named Entity Recognition und Information Retrieval auf Wirtschaftsnachrichten sind in den letzten Jahren immer beliebter geworden, und wenn man an das frei verfügbare Reuters Korpus[28] denkt, das für Studien dieser Art sogar noch aufbereitet wurde, stellt man fest, dass der Bedarf an Informationsextraktion aus wirtschaftlich orientierten Texten bei weitem noch nicht gedeckt ist. Mit der immer stärker werdenden Verflechtung internationaler Wirtschaftsbeziehungen, dem ständig anwachsenden Trend der internationalen Fusionen und der Globalisierung der Wirtschaft wächst die Nachfrage aus aktuellsten Wirtschaftsartikeln, kurz und prägnant interessante Information zu erhalten. Der Kreis der Suchenden beschränkt sich heute längst nicht mehr nur auf Betriebs- oder Volkswirte, sondern auf jeden, der in die Wirtschaft investieren möchte, und sich aufgrunddessen informiert.

All diese Gründe machen Wirtschaftsnachrichten zu einer lukrativen und begehrten Domäne für die Informationsgewinnung und heben die Nachfrage nach qualitativ guten Systemen zur Wissensextraktion auf Nachrichtentexten.

[28] `http://about.reuters.com/researchandstandards/corpus/`

4.3 Priorisierung von Entitäten

In Kapitel 3 wurden bereits unterschiedliche Ansätze zur Erkennung benannter Entitäten (Named Entities) mittels lokaler Grammatiken vorgestellt. All diese Ansätze haben nicht nur die Gemeinsamkeit, dass sie sprachbasierte statt statistische Methoden zur Lokalisierung von Eigennamen oder Verbgefügen anwenden, sondern auch dass keiner dieser Linguisten versucht hat, alle Kategorien von Entitäten in einem System zur Named Entity Recognition zusammenzufassen. Jeder von ihnen hat sich auf eine Entität konzentriert – einige auf Personen und andere auf Organisationen. Natürlich spielten immer wieder andere Entitäten, wie vorallem Toponyme, eine untergeordnete Rolle bei der Erkennung von Menschen oder Firmen. Meist waren sie dann nur Mittel zum Zweck, indem sie Teil des Kontextes der zu suchenden Entität waren.

Personen werden stets eine der beliebtesten Entitäten für die NER sein, auch wenn die automatische Produktnamenerkennung inzwischen immer mehr in den Vordergrund rückt, wie es die Arbeit von Jeannette Roth [Roth, 2002] zeigt. Bis jetzt werden wohl die syntaktischen und semantischen Aspekte von Produktnamen noch unerforschter sein als die Eigenschaften von Organisationsnamen. Dennoch bewies auch die Arbeit von Friederike Mallchok [Mallchok, 2004], wie gut lokale Grammatiken das Problem der Organisationsnamenerkennung lösen können.

Gerade bei der Suche auf Wirtschaftsnachrichtentexten stößt man auf eine beträchtliche Anzahl von Organisationsnamen. Diese Kategorie der verschiedenen Entitäten wird jedoch in dem hier präsentierten Ansatz eine untergeordnete Rolle zu den Menschenbezeichnern haben. Da aber Personen in Wirtschaftsartikeln sehr häufig im Zusammenhang mit Firmen genannt werden und ihr Verhältnis zu diesen oft explizit beschrieben wird, sollte natürlich den Beziehungen zwischen diesen beiden Entitäten besonders viel Beachtung geschenkt werden. Obwohl diese beiden Gruppen – Personen und Organisationen – die frequentesten Entitäten in Wirtschaftstexten sein werden, gibt es dort noch viele weitere personenbezogene Relationen, in deren Kontext womöglich andere Entitäten wie Ortsbezeichnungen auftreten können.

Für alle Entitäten, die keine Menschenbezeichner sind, werden lokale Grammatiken erstellt, welche dazu dienen, das Umfeld der Personen zu spezifizieren. Die Grammatiken entsprechen in ihrem Umfang und in ihrer Ausführlichkeit der Wichtigkeit der Relation, die zwischen der jeweiligen Entität und der Personenbezeichnung herrscht. Somit werden die Grammatiken für die Organisationsbezeichner umfassender als für die Toponyme sein, da sie eine größere Relevanz in Bezug auf das Korpus haben.

Für diesen Ansatz gilt, dass die Menschenbezeichner als Entität priorisiert werden. Doch würde es für zukünftige Vorhaben kein Problem darstellen, die Gewichtung der Entitäten für die jeweiligen Zwecke abzuändern.

5 Ressourcen: Grundlagen des Systems

5.1 Korpora

Wie bereits mehrfach erwähnt wurde, sollten Menschenbezeichner innerhalb biographischer Relationen automatisch in Wirtschaftsnachrichten erkannt werden. Diese Vorgabe schränkt die Wahl der Texte, auf denen gearbeitet werden kann, zunächst auf die Wirtschaftsteile vieler englischsprachiger Zeitungen ein. Nur wer begnügt sich mit dem Wirtschaftsteil, wenn ganze Wirtschaftsblätter ihre Artikel online zur Verfügung stellen?

Ähnlich wie bei Friederike Mallchok [Mallchok, 2004] wäre das Reuters Korpus eine Option gewesen, da es eine Textsammlung aus Wirtschaftsartikeln ist. Doch die Tatsache, dass es für Wirtschaftsnachrichten relativ veraltet ist, machte es zu keinem Kandidaten für ein Testkorpus.

Dagegen war das Angebot vom Centrum für Informations- und Sprachverarbeitung der LMU München, mir eine Jahresausgabe der Financial Times (FT) zur Verfügung zu stellen, wesentlich interessanter. Vorallem handelte sich hierbei um die Jahresausgabe 2004 der FT, womit sicher gestellt ist, dass die darin enthaltenen Informationen relativ aktuell sind.

5.1.1 Financial Times

Die Financial Times[29] ist eine Tageszeitung, welche fast täglich herausgegeben wird. In ihrem elektronischen Format ist jede Tagesausgabe eine XML-Datei und das Jahr 2004 umfasste 347 Tage, an denen die FT erschienen ist. Somit ergab sich eine Datenmenge von ungefähr 5,8 GB.

Um aus dieser Artikelsammlung ein Korpus zu erstellen, wurden zunächst alle Texte von ihrer XML-Information befreit, was die Größe der Daten auf 4,7 GB verminderte. Im Anschluss wurden die Tagesausgaben monatsweise zusammengefügt, so dass es für jeden Monat eine Datei der Financial Times gab.

Diese 12 Dateien wurden nun für die spätere Bearbeitung mit dem System UNITEX [Paumier, 2004] vorbereitet:

- Im ersten Schritt wurde die Satzenderkennung mit dem Tokenizer-Programm[30] von Sebastian Nagel auf dem gesamten Text vorgenommen.

  Ein Programmaufruf folgender Form

  ```
  cat <korpus> | tokenizer -L en -SE {S} -P -o <korpus.eos>
  ```

[29] `http://news.ft.com/home/us`

[30] Eine aktuelle Version des Tokenizer-Programms ist unter `http://www.cis.uni-muenchen.de/~wastl/misc/tokenizer.tgz` verfügbar.

liefert einen Text mit Satzendmarkierungen, wie ihn das Programm UNITEX fordert. Mir wurde dieses Programm in der Version 0.6 überlassen, so dass es für die Satzenderkennung im Englischen angepasst werden konnte, da die deutsche Satzenderkennung ausgereifter als die englische war. Diese Verbesserungen wurden anschließend in die Version 0.7 aufgenommen.

Das Programm UNITEX bietet zwar auch eine Satzenderkennung für das Englische an, doch handelt es sich bis jetzt um die französische Satzenderkennung, die nur leicht für das Englische abgewandelt wurde und leider immer noch größtenteils die französischen Abkürzungen enthält. Somit war diese Satzenderkennung keine Alternative zum Tokenizer-Programm, was wirklich hervorragende Ergebnisse geliefert hat.

- Im nächsten Schritt wurde die Normalisierung und Tokenisierung des Textes mit den entsprechenden Programmen aus dem System UNITEX vorgenommen.

- Im letzten Schritt wurde die gesamte grammatikalische und semantische Information aus den im nächsten Abschnitt angesprochenen Lexika im Korpus passend annotiert. Das heißt aber nicht, dass der Originaltext verändert wurde, sondern dass diese Zusatzinformationen in Wortlisten ergänzend zum Text gespeichert werden.

Somit ist das FT-Korpus für die Entwicklung lokaler Grammatiken mit den System UNITEX bereit, welche anschließend darauf getestet werden können.

5.1.2 Biography.com

Dennoch ist das FT-Korpus nicht die einzige Textsammlung, welche zur Validierung der hier vorgestellten lokalen Grammatiken verwendet wird.

Wie bereits in Kapitel 4 angedeutet wurde, ist nicht jede biographische Relation in Wirtschaftsnachrichten wie der Financial Times vertreten. Manche von ihnen sind typisch für Lebensläufe und könnten in einem Korpus basierend auf diversen Biographien sicher gefunden werden.

Aufgrund dessen wurde ein Crawler implementiert, der alle fast 25.000 Biographien der Internetseite *Biography.com*[31] heruntergeladen hat. So standen nochmal knapp 100 MB Daten zum Testen der Grammatiken zur Verfügung.

Hierbei handelte es sich um HTML-Dateien, welche zunächst in reinen Text umgewandelt wurden und daraufhin durchliefen sie die gleichen Bearbeitungsschritte wie das Financial Times Korpus, so dass das Biography.com-Korpus im System UNITEX verwendet werden konnte.

[31] `http://www.biography.com`

5.2 Wörterbuchressourcen

Zur Verarbeitung dieser Korpora, wurden mehrere Lexika erstellt. Diese Wörterbücher könnten mit einigen Anpassungen in ihrem Format später auch ohne weiteres in das CISLEX-E [Guenthner, 1996] integriert werden. Das CISLEX-E ist ein am Centrum für Informations- und Sprachverarbeitung entwickeltes morphologisches, syntaktisches und semantisches Lexikon der englischen Sprache.

Natürlich wird die Erkennung von Menschenbezeichnern in biographischen Kontexten um einiges einfacher, wenn das System später auf eine relativ große Wissensbasis zurückgreifen kann. Durch qualitativ gute Lexikoneinträge lässt sich auch die Performanz und Genauigkeit des Systems deutlich verbessern, da Ambiguitäten leichter aufgelöst werden können, ohne dass der Kontext miteinbezogen werden muss. Das soll nichts anderes bedeuten, als dass ein im Text auftretender Personenname als dieser eindeutig identifiziert werden kann, falls er bereits im Wörterbuch vorkommt. Wenn der Name also nicht in einem der Lexika kodiert wurde, muss man sich an bestimmten Kontextschemata orientieren, um herauszufinden, ob es sich hierbei um einen Personennamen handelt. Gibt es keinen eindeutigen Kontext, der darauf schließen lässt, wird das Auffinden dieses Namens fast unmöglich.

Muss man sich jedoch an den Kontexten, in denen ein Menschenbezeichner eingebettet sein kann, orientieren, so ist es durchaus hilfreich mehr Entitäten als nur Personennamen zu sammeln und in Lexika festzuhalten.

Ein Personenname kann beispielsweise aus einem Titel oder einer Anrede gefolgt von einem Nachnamen bestehen. So wäre es nur sinnvoll, Wörterbücher allein nur für Titel und Anredemöglichkeiten zu erstellen, sowie Vor- und Nachnamen gesondert aufzulisten, aber auch vollständige Personennamen zu archivieren.

Diese Hyperonymierelationen können auch für die allgemeinen Menschenbezeichner definiert werden, um festzulegen, welche englischen Nomina auf einen Menschen referenzieren, um später eine Übergenerierung der Graphen zu verhindern. Dadurch schränken wir die grammatikalische Klasse der Nomina auf diesen Teil ein und verhindern somit, dass später beispielsweise Tiere statt Menschen im Text gefunden werden.

Da in biographischen Texten häufig Beschäftigungsverhältnisse beschrieben werden, sollte man auch nicht auf die Kategorie der Berufsbezeichnungen verzichten. In diesem Zusammenhang sind Organisationsnamen bzw. Firmennamen und organisationsspezifische Attribute nicht außer Acht zu lassen. Auch eine Liste an Branchen, Fachbereichen und Industriesektoren kann von Vorteil sein, wenn nur die Arbeitsdomäne einer Person genannt wird.

Des Weiteren kommen in diesen Kontexten häufig Ortsbestimmungen und Beschäftigungszeiträume vor, was wiederum den Einsatz eines Wörterbuches mit geographischen Bezeichnungen nötig macht. Aber auch für biographisch typische Sätze wie *„Harry Clifford was born in Dallas, Texas in 1956.“* sind Toponyme ein unverzichtbares Muss.

All diese Vorüberlegungen zeigen, welche weiteren Entitäten außer Personenbezeichnungen für diese Aufgabe notwendig sein werden. Deshalb habe ich verschiedene Lexika mithilfe von Listen aus dem Internet erstellt, welche dem System eine riesige Wissensbasis liefern werden. Insgesamt haben alle Wörterbücher zusammen mehr als 10,6 Millionen Einträge.

5.2.1 Lexikon der Vornamen

Das Lexikon der Vornamen enthält weit über 38.500 Einträge und hat den Namen *FirstNames-.dic*. Die darin enthaltenen Vornamen stammen aus dem CISLEX, sowie aus verschiedenen Listen des Internets ([Wikipedia, 2005/2006] ist nur eine von vielen Quellen.). Es wurden einfache Vornamen wie *„Mary“*, aber auch komplexe Vornamen wie *„Mary-Anne“* aufgelistet.

In Abbildung 5.1 wird deutlich, in welcher Form die Vornamen gespeichert werden. An dieser Stelle soll für uns das Format der Kodierung nicht von Interesse sein, aber die Bedeutung der Symbole nach dem jeweiligen Vornamen möchte ich schon einmal hier erläutern. Es ist möglich bei der Erstellung der Lexika anzugeben, welchen semantischen oder grammatikalischen Kategorien die jeweiligen Einträge angehören. Außerdem kann man eigene Kategorien selbst definieren und die Wörter nach diesen klassifizieren. Für *FirstNames-.dic* wurden 4 Kategorien verwendet, von denen die Klassen `N`, `PR` und `Hum` dem System UNITEX [Paumier, 2004] bekannte Kategorien für die englische Sprache sind, und `FN` wurde von mir eingeführt. Die Tabelle 5.1 erläutert die Bedeutung dieser Abkürzungen.

```
Anneliese,.N+PR+Hum+FN
Barbara-Anne,.N+PR+Hum+FN
Charly,.N+PR+Hum+FN
Delores,.N+PR+Hum+FN
Elizabeth,.N+PR+Hum+FN
Frédérique,.N+PR+Hum+FN
```

Abbildung 5.1: Auszug aus dem Lexikon *FirstNames-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
PR	semantisch	Proper Noun	Eigenname
Hum	semantisch	Human	Menschenbezeichner
FN	semantisch	First Name	Vorname

Tabelle 5.1: Abkürzungen, die in *FirstNames-.dic* verwendet werden.

5.2.2 Lexikon der Nachnamen

Das Lexikon der Nachnamen enthält ca. 1,25 Millionen Einträge und trägt den Namen *LastNames-.dic*. Die darin enthaltenen Nachnamen stammen wie schon bei dem Lexikon der Vornamen aus dem CISLEX, sowie aus verschiedenen Listen aus dem Web. Auch hier wurden einfache Nachnamen wie *„Smith“*, aber auch komplexe Nachnamen wie *„Amesöder-Gerogan“* gesammelt.

Abbildung 5.2 zeigt einige Einträge aus dem Wörterbuch der Nachnamen, wofür ich auch 4 Kategorien verwendet habe, von denen wieder die Klassen `N`, `PR` und `Hum` dem System UNITEX [Paumier, 2004] bekannte Kategorien für die englische Sprache sind und `SN` wurde von mir eingeführt. Die Tabelle 5.2 erläutert die Bedeutung dieser Abkürzungen.

```
Aabenhus,.N+PR+Hum+SN
Amesöder-Gerogan,.N+PR+Hum+SN
Gabanelli,.N+PR+Hum+SN
MacMillan,.N+PR+Hum+SN
Olsson,.N+PR+Hum+SN
Töchert-Yildiz,.N+PR+Hum+SN
```

Abbildung 5.2: Auszug aus dem Lexikon *LastNames-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
PR	semantisch	Proper Noun	Eigenname
Hum	semantisch	Human	Menschenbezeichner
SN	semantisch	Surname	Nachname

Tabelle 5.2: Abkürzungen, die in *LastNames-.dic* verwendet werden.

5.2.3 Lexika der Personennamen

Die Lexika der Personennamen enthalten zusammen ungefähr 8,3 Millionen Einträge.

Wie in Tabelle 5.3 auf Seite 50 ersichtlich wird, wurden auch hier für die Kodierung der Wörterbucheinträge 4 Kategorien verwendet, von denen wieder die Klassen `N`, `PR` und `Hum` dem System UNITEX [Paumier, 2004] bekannte Kategorien für die englische Sprache sind und `LN` wurde hier eingeführt. Leider konnte nicht die Abkürzung `FN` für *Full Name* gewählt werden, da `FN` schon als Kategorie für *First Name* bzw. Vornamen im *FirstNames-.dic* belegt war, musste auf `LN` für *Long Name* ausgewichen werden.

Personennamen von Biography.com

Das Internetportal *Biography.com* [Biography.com, 2005/2006] stellt knapp 25.000 englischsprachige Biographien zur Verfügung. Jede einzelne Biographie enthält in der Regel einen Personennamen. Somit bietet Biography.com indirekt fast 25.000 Personennamen, bestehend aus Vor- und Nachnamen im Web an.
Variiert man nun die Stellung von Vor- und Nachname, so dass

```
Aaron\, Hank,.N+PR+Hum+LN
Hank Aaron,.N+PR+Hum+LN
```

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
PR	semantisch	Proper Noun	Eigenname
Hum	semantisch	Human	Menschenbezeichner
LN	semantisch	Long Name (Full Name)	Vollständiger Personenname, bestehend aus Vor- und Nachname, evtl. mit Titel

Tabelle 5.3: Abkürzungen, die in *LongNames*-.dic* verwendet werden.

ins Wörterbuch aufgenommen werden, so verdoppeln sich die Einträge.

Auch mit der Abkürzung zweiter Vornamen, lassen sich weitere Lexikoneinträge gewinnen, wie z.B.

```
Aaron T\. Beck,.N+PR+Hum+LN
Aaron T(emkin) Beck,.N+PR+Hum+LN
```

Außerdem können selten benutzte zweite Vornamen auch weggelassen oder geklammert werden, was zwei neuen Einträge entstehen lässt:

```
Aage Bohr,.N+PR+Hum+LN
Aage (Niels) Bohr,.N+PR+Hum+LN
Aage Niels Bohr,.N+PR+Hum+LN
```

Mithilfe dieser Methoden lassen sich aus knapp 25.000 Personennamen, mehr als 69.000 Wörterbucheinträge generieren und im Lexikon *LongNamesBios-.dic* speichern.

Personennamen aus der Financial Times

Natürlich bietet das aus der Financial Times[32] zusammengestellte Korpus eine Fülle an Personennamen.

Um auf diese Namen nicht verzichten zu müssen, wurde zunächst nach zwei nebeneinander groß geschriebenen Wörtern, in deren Mitte evtl. noch ein einzelner Großbuchstabe auftreten darf, in dem gesamten Text gesucht.
Diese Suche würde u.a.

```
George W Bush
Henna Nordqvist
```

als Treffer in den Zeitungsberichten ergeben.

Da in englischen Texten in der Regel nur Eigennamen groß geschrieben werden, wenn der Satzanfang außer Acht gelassen wird, kann man mit großer Wahrscheinlichkeit davon ausgehen, dass alle Eigennamen im Korpus auf diese Weise gefunden werden.

[32] `http://news.ft.com/home/us`

Diese Eigennamen sind sicher nicht alle nur Personennamen, doch mehr als die Hälfte von ihnen fallen in diese Kategorie. Um Organisationsnamen wie *„Peerless Industries“* aussortieren zu können, musste die Liste der potentiellen Personen von mir manuell durchgesehen werden. So konnten Firmennamen oder Filmtitel letztendlich aus diesem Lexikon entfernt werden.

Insgesamt blieben ca. 630.000 richtige Personennamen von ursprünglich 701.778 Kandidaten übrig, welche im Wörterbuch mit dem Namen *LongNamesFT-.dic* gespeichert wurden.

Dieses Ergebnis lässt nicht darauf schließen, dass knapp 7% von dieser Eigennamenliste Organisationsnamen gewesen sein müssen. Einerseits wären das viel zu wenig Firmennamen für eine Jahresausgabe der Financial Times, und andererseits wurden vor der manuellen Nachkorrektur, alle Organisationsnamen, welche durch ihre Rechtsform (z.B. GmbH, AG, S.A., PLC, Inc., etc.) gekennzeichnet waren, aus dieser Kandidatenliste herausgenommen und gesondert aufbewahrt.

Personennamen von SpecialistInfo.com

SpecialistInfo.com[33] bietet auf seiner Seite u.a. eine Liste von Personen, die als Berater tätig sind (Consultants). Diese enthält derzeit 31.000 Personennamen, welche mit ihrem jeweiligen Titel (akademischer Grad, militärischer Rang usw.) oder der jeweiligen Anredeform (*„Mr“* oder *„Mrs“*) versehen sind.

Hier hätte man auch die Möglichkeit gehabt, durch Entfernen der Titel und Anredeformen – welche eigentlich nur den Beruf bzw. das Geschlecht der Person Aufschluss geben – die Anzahl der späteren Lexikoneinträge zu verdoppeln. Doch da die meisten dieser Namen ohne Titel oder Anrede, in bereits genannten oder später genannten Lexika namentlich vorkommen, entschied ich mich dazu, auf diesen Schritt zu verzichten. Somit hat das Wörterbuch *LongNamesSpecialistInfo-.dic* nahezu 31.000 Einträge.

Personennamen von ZoomInfo.com

ZoomInfo.com[34] ist eine Suchmaschine für Personen- und Firmennamen. Da sie über einen riesigen Datenbestand verfügt, konnten ca. 3,5 Millionen Personennamen extrahiert werden.

Weitere Personennamen

Wie bereits Latanya Sweeney in ihrem Artikel *„Finding Lists of People on the Web“* [Sweeney, 2004] so treffend bemerkt hatte, kann man auch durch Zufall auf riesige Listen mit Personennamen im Internet stoßen. Ähnlich erging es mir bei einer Personenrecherche, als mir Google zwei Listen mit Autoren[35] anbot. So erhielt ich für das Lexikon *LongNamesAuthors-.dic* insgesamt 449.500 Einträge.

[33] `http://www.specialistinfo.com/directory.php`
[34] `http://www.zoominfo.com`
[35] `http://136.199.54.185/~ley/db/indices/AUTHORS`
`http://sunsite.online.globule.org/dblp/db/indices/AUTHORS`

5.2.4 Lexika der Personentitel

Für den Fall dass ein Personenname im Text auftaucht, welcher nicht in einem der Wörterbücher registriert ist, muss man den Kontext zur Erkennung des Namens miteinbeziehen. Ein eindeutiger Indikator für einen Personennamen ist beispielsweise ein akademischer Grad, ein militärischer Rang, ein aristokratischer Titel oder einfach nur eine Form der Anrede. Aus diesem Grund habe ich mit der Hilfe der englischen Wikipedia [Wikipedia, 2005/2006] ein Wörterbuch von verschiedenen Titelbezeichnungen zusammengestellt. Dieses trägt den Namen *Titles-.dic* und enthält über 370 Einträge.

```
Duke,.N+Title
Mrs,.Abbrev+Title
Rabbi,.N+Title
Sir,.N+Title
Sister,.N+Title
```

Abbildung 5.3: Auszug aus dem Lexikon *Titles-.dic*

Akademische Grade

```
DUniv,.Abbrev+Title
MSci,.Abbrev+Title
Prof\. h\.c\.,.Abbrev+Title
Prof\.,.Abbrev+Title
PsyD,.Abbrev+Title
```

Titles-.dic enthält eine Fülle an Abkürzungen für akademische Grade. Meist besteht im englischen Sprachgebrauch die Möglichkeit diese Titel mit nachfolgendem Punkt oder ohne diesen zu schreiben. Um nun alle Möglichkeiten, wie eine Titelbezeichnung im Text auftreten kann, abzudecken, wurden auch diese Variationen ins Lexikon aufgenommen.

Aristokratische Titel

```
Duke,.N+Title
Earl,.N+Title
Lord,.N+Title
Queen,.N+Title
Prince,.N+Title
Princess,.N+Title
```

Oft wird auch in den Nachrichten über adelige Persönlichkeiten berichtet. Deren Namen wird ein Adelstitel vorangestellt. Darum ist es ebenfalls wichtig, die Gruppe dieser Titel in das Wörterbuch aufzunehmen. So können später auch unbekannte Adelige, die irgendeiner Nebenlinie entstammen, und deshalb in keinem der Personenlexika aufgeführt sind, im Text gefunden werden. Auch werden bei Adelstiteln im Gegensatz zu akademischen Graden in der Regel keine Abkürzungen verwendet.

Weitere Anredeformen

```
Captain,.N+Title
Mr,.Abbrev+Title
Patriarch,.N+Title
People's Commissar,.N+XN+Title
Pope,.N+Title
Saint,.N+Title
```

Es gibt weitere zahlreiche Anredemöglichkeiten oder Titel für die unterschiedlichsten Personengruppen, seien es jetzt Vertreter der Kirche, des Militärs oder einfach nur die Anredeform *„Mr“*, *„Mrs“* oder *„Ms“*. Da keine dieser Personengruppen dominanter als die andere in diesem Wörterbuch ist, werden sie in diesem Abschnitt zusammen erwähnt.

5.2.5 Lexika der allgemeinen Menschenbezeichner

Bis zu diesem Punkt wurden die Lexika der Personeneigennamen beschrieben, die sicherlich auch zu den Menschenbezeichnern zählen. Doch in diesem Abschnitt soll es nun ausschließlich um Bezeichnungen für Menschen gehen, die keine Eigennamen sind. Unter *allgemeinen Menschenbezeichnern* versteht man beispielsweise Begriffe, die Verwandtschaftsverhältnisse wie *„mother“* ausdrücken, Berufsbezeichnungen wie *„salesperson“* oder Personen durch ihren Charakter wie *„idealist“* bzw. ihre Neigungen wie *„lesbian“* näher bestimmen. Des Weiteren fallen auch Wörter, welche eine Staatszugehörigkeit oder die Zugehörigkeit zu einem Bezirk, einer Provinz, einer Stadt oder eines Stadtteils ausdrücken, in diese Kategorie der allgemeinen Menschenbezeichner. Aus diversen Internetverzeichnissen habe ich insgesamt über 54.000 Begriffe zusammengetragen, die auf diese Beschreibungen zutreffen. Aus einer früheren Arbeit von Friederike Mallchok [Mallchok, 2004] konnte ich auf ein Lexikon mit mehr als 99.000 Einträgen für Berufsbezeichnungen zurückgreifen, so dass mir ein Inventar von insgesamt ca. 153.000 allgemeinen Menschenbezeichnern zur Verfügung stand.

Allgemeine Menschenbezeichnungen aus WordNet

WordNet [WordNet, Version 2.1] ist eine echte Bereicherung, wenn man Begriffe mithilfe semantischer Relationen wie der Hyperonymie sucht. WordNet [Fellbaum, 1998] bezeichnet sich selbst als elektronische lexikalische Datenbasis. Im Grunde teilt WordNet das Lexikon in fünf Kategorien auf: Nomen, Verben, Adjektive, Adverbien und Funktionswörter [Miller *et al.*, 1993]. Dabei verfolgt WordNet die Zielsetzung lexikalische Information mehr durch Wortbedeutungen zu organisieren als durch Wortformen. Die Lexikoneinträge für Nomen sind nicht alphabetisch sondern hierarchisch geordnet.

```
{ Nomen-Oberbegriff (semantische Merkmale) }
```

Dabei werden die Nomen in Synsets klassifiziert, wobei ein Synset die Menge alle Synonyme eines Nomens ist. Es wurden alle wichtigen semantischen Relationen wie Synonymie, Meronymie, Holonymie, Hyperonymie, Hyponymie und Antonymie kodiert, und man

könnte die semantische Relationen von WordNet als Relationen zwischen den einzelnen Synsets beschreiben. Außerdem ist die große Masse von Nomen hierarchisch in 25 eindeutigen initialen Synsets (*25 unique beginners*) organisiert[36].

Da Menschenbezeichner in der Regel als Nomen vorkommen, möchte ich nicht näher auf die anderen grammatikalischen Kategorien eingehen, welche in WordNet noch vertreten sind. Außerdem ist für uns momentan nur das Anfangssynset `{person, human being}` von Bedeutung, in dem alle Menschenbezeichner in WordNet gespeichert sind.
Die Gruppe der Menschenbezeichner aus WordNet enthält alle oben genannten Variationen an allgemeinen Menschenbezeichnungen und darüber hinaus noch ganz allgemeine Berufsbezeichnungen wie *„engineer“* oder *„worker“*. Das mithilfe von WordNet erstellte Wörterbuch *MenbezWordnet-.dic* kann insgesamt ungefähr 6.400 Einträge aufweisen.

```
godmother,.N+Hum
heterosexual,.N+Hum
husband,.N+Hum
idealist,.N+Hum
islander,.N+Hum
```

Abbildung 5.4: Auszug aus dem Lexikon *MenbezWordnet-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
Hum	semantisch	Human	Menschenbezeichner

Tabelle 5.4: Abkürzungen, die in *MenbezWordnet-.dic* verwendet werden.

Berufsbezeichnungen

Das Lexikon der Berufsbezeichnungen ist aus einer Vielzahl von Internetverzeichnissen[37], den Listen von Job-Suchmaschinen[38] oder von Arbeitsämtern[39] aus dem englischsprachigen Raum entstanden.

Auch wurde ich noch bei der englischen Wikipedia [Wikipedia, 2005/2006] auf meiner Suche nach Berufsbezeichnern fündig und stellte so ein Wörterbuch von verschiedenen Berufsbezeichnungen zusammen. Dieses trägt den Namen *JD-.dic* und enthält nahezu 47.000 Einträge.

[36] vgl. [Geierhos, 2004]

[37] Guide to the World of Occupations [Guide to the World of Occupations, 2005], Occupational Outlook Handbook [OOH, 1998] [OOH, 2000/2001], List of occupations [Lacombe, 2004], Standard Occupational Classification (SOC) [UBC, 1991], Dictionary Of Occupational Titles [DOT, 2005]

[38] CareerBuilder.com [CareerBuilder.com, 2005], LabourMarket [LabourMarket, 2005], Prospects.ac.uk [Prospects.ac.uk, 2005]

[39] Canadian Job Classification [Government of Newfoundland und Labrador, 2005], Division of Professional Licensure [DoPL, 2005], Ministry of Manpower [MoM, 2003], U.S. Department of Labor [U. S. Department of Labor, 2005] [USDoL, 2001/2002]

```
accountant,.N+Hum+JD
adjuster and inspector,.N+XN+Hum+JD
aeronautical technician,.N+XN+Hum+JD
bakery products checker,.N+XN+Hum+JD
bank president,.N+XN+Hum+JD
blacksmith,.N+Hum+JD
car salesman,.N+XN+Hum+JD
```

Abbildung 5.5: Auszug aus dem Lexikon *JD-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
XN	grammatikalisch	Extended Noun	Mehrwortlexem
Hum	semantisch	Human	Menschenbezeichner
JD	semantisch	Job Descriptor	Berufsbezeichner

Tabelle 5.5: Abkürzungen, die in *JD-.dic* verwendet werden.

Einwohnerbezeichnungen

Eine relativ große Gruppe innerhalb der Menschenbezeichner, sind die Bezeichnungen nach Herkunftsland oder -ort. Bei der Zusammenstellung dieser Liste habe ich mich ganz darauf konzentriert möglichst alle Menschenbezeichner, die sich auf Staatsbürgerschaften beziehen, in das neue Lexikon aufzunehmen. Bei den Begriffen, welche ausdrücken, dass jemand Einwohner einer Stadt, eines Ortes, eines Bezirks, einer Provinz, eines Bundesstaates oder Bundeslandes ist, habe ich mich auf die wichtigsten und bekanntesten Regionen und Städte beschränkt, da Bezeichnungen wie *„New Yorker“* im Korpus häufiger als z.B. *„Nottinghamian“* auftreten.

Auch hier wurde ich bei der englischen Wikipedia [Wikipedia, 2005/2006] fündig und stellte so ein Wörterbuch zusammen, was den Namen *Citizens-.dic* hat und über 600 Einträge enthält.

```
Albertans,Albertan.N+Citizen+CaProvinceCitizen+Hum
Americans,American.N+Citizen+Hum
Athenians,Athenian.N+Citizen+Urbanite+Hum
Brooklyners,Brooklyner.N+Citizen+Urbanite+NYCcitizen+Hum
Bavarians,Bavarian.N+Citizen+Hum
Canberrans,Canberran.N+Citizen+AuProvinceCitizen+Hum
Christmas Islanders,Christmas Islander.N+Citizen+Hum
Ontarians,Ontarian.N+Citizen+CaProvinceCitizen+Hum
Tasmanians,Tasmanian.N+Citizen+AuProvinceCitizen+Hum
```

Abbildung 5.6: Auszug aus dem Lexikon *Citizens-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
Hum	semantisch	Human	Menschenbezeichner
Citizen	semantisch	Citizen	(Staats)Bürger
AuProvinceCitizen	semantisch	Australian Province Citizen	Bewohner einer australischen Provinz
CaProvinceCitizen	semantisch	Canadian Province Citizen	Bewohner einer kanadischen Provinz
USstateCitizen	semantisch	US State Citizen	Bewohner eines US Bundesstaates
NYCcitizen	semantisch	New York City Citizen	Bewohner von New York City
Urbanite	semantisch	Urbanite	Stadtbewohner

Tabelle 5.6: Abkürzungen, die in *Citizens-.dic* verwendet werden.

5.2.6 Lexikon der personenbezogenen Prädikate

In Abschnitt 1.2 wurde bereits der Begriff der biographischen Relation erläutert. Allen drei Kategorien von Relationen – der persönlichen, der öffentlichen und der zufälligen Relation – lassen sich Verben bzw. Verbalphrasen der englischen Sprache zuordnen. Obwohl anfangs die Priorisierung von öffentlichen Relationen in den Vordergrund gestellt wurde, soll für die Wörterbucherstellung trotz möglicher Überschneidungen keine Differenzierung zwischen den drei Verbkategorien vorgenommen werden.

Für die Erkennung von Menschenbezeichnern in biographischen Kontexten ist die Einschränkung der potentiellen Verben, welche in der Umgebung von Personen vorkommen, fast unumgänglich. Deshalb müssen Prädikate (insbesondere Verben) ausgewählt werden, welche zumindest an erster Argumentposition vornehmlich einen Menschenbezeichner fordern. Diese personenbezogenen Prädikate können natürlich auch als zweites Argument eine Person verlangen, doch reicht die erste Einschränkung in der Regel aus.

Um nun ein solches Lexikon erstellen zu können, hat man die Möglichkeit, alle relevanten Verben der englischen Sprache aufzulisten, oder die vorhandenen Ressourcen werden mit in die Erstellung eines personenbezogenen Verbalphrasenlexikons einbezogen. Für die weitere Vorgehensweise wurden alle Personennamen des Financial Times Korpus annotiert und die Menge der Sätze, die einen Namen enthalten, extrahiert. Danach wurden die lokalen Grammatiken zur Lemmatisierung komplexer englischer Verben (siehe Abschnitt 3.2, vgl. [Gross, 1998–1999]) auf den neuen Text angewendet und die entsprechenden Treffer markiert. Im Anschluss wurden alle erkannten Verbalphrasen, denen eine Nominalphrase mit einem Menschenbezeichner vorangestellt war, dem Korpus entnommen und noch einmal manuell überprüft, bevor man sie dem neuen Lexion *HumVP-.dic* hinzugefügt hat. Die noch fehlenden einfachen Verben wurden aus den zusammengesetzten Verbeinträgen rekonstruiert.

```
accused of committing,.HumVP
agreed to marry,.HumVP
are resigning,.HumVP
attempted to establish,.HumVP
can join,.HumVP
cannot risk losing,.HumVP
could be fired,.HumV
...
```

Abbildung 5.7: Auszug aus dem Lexikon *HumVP-.dic*

5.2.7 Lexika der Branchen

Auch Branchen treten recht häufig im Umfeld von Menschenbezeichnern bzw. Berufsbezeichnern auf. Meist dienen sie als Ergänzung oder Spezifikation eines Arbeitsbereichs. Diese Branchenbezeichnungen kann man aufgrund ihres syntaktischen Verhaltens in zwei Bereiche aufteilen: Fachbereiche und Lehrfächer, sowie Sektoren- und Branchenbezeichnungen.

Fachbereiche und Lehrfächer

Die Fachbereiche und Lehrfächer zeichnet besonders aus, dass sie in der Regel im rechten Kontext von Berufsbezeichnern zu finden sind, welche akademische Berufe, Lehrtätigkeiten oder Spezialisierung von Berufen ausdrücken. Des Weiteren fordern die jeweiligen Fachbereiche selbst keine Ergänzungen für sich selbst, was nichts anderes heißt, als dass beispielsweise *american literature* in der Nominalphrase *„professor of american literature“* kein eigenes Argument mehr braucht – es dient selbst als Ergänzung zu *professor*.

Die gesammelten Wörterbucheinträge ließen sich aus Seiten der englischen Wikipedia [Wikipedia, 2005/2006] extrahieren und ergaben ein Begriffsinventar von ca. 580 Einträgen für das Lexikon *Disciplines-.dic*.

```
american literature,.N+XN+Sector+Discipline
analytical chemistry,.N+XN+Sector+Discipline
bioinformatics,.N+Sector+Discipline
cognitive science,.N+XN+Sector+Discipline
descriptive linguistics,.N+XN+Sector+Discipline
```

Abbildung 5.8: Auszug aus dem Lexikon *Disciplines-.dic*

Sektoren- und Branchenbezeichnungen

Ganz anders verhalten sich dagegen die Sektoren- und Branchenbezeichnungen. Sie fordern sehr wohl eine Ergänzung in Form von Begriffen wie *„industry“*, *„sector“* oder *„company“* und bestimmen somit genauer, in welcher Branche eine Person beschäftigt ist.

Somit sind im rechten Kontext dieser Branchenbegriffe, die unterschiedlichsten Schlüsselbegriffe zu finden.

```
administration sector
animal food industry
arts and leisure sector
```

Die eben genannten Sektoren könnten allerdings auch mit dem Schlüsselbegriff *„services“* versehen werden. Daraus wird ersichtlich, dass es eine Gruppe von Branchenbezeichnungen gibt, die flexibler als andere in der Wahl ihrer Ergänzungen sind. Es gibt eine Reihe von Begriffen, welche nur mit der Ergänzung *„services“* im Text auftreten. Es wäre beispielsweise unsinnig *„animal physiotherapy services“* ein anderes Argument als *„services“* mitzugeben. Solche Branchenbegriffe, die vornehmlich oder ausschließlich mit der Ergänzung *„services“* in Texten vorkommen, wurde die Ergänzung bei der Wörterbucherstellung mitgegeben.

```
accident and health insurance,.N+XN+Sector
administration,.N+Sector
aircraft part and auxiliary equipment,.N+XN+Sector
animal food,.N+XN+Sector
arts and leisure,.N+XN+Sector
```

Abbildung 5.9: Auszug aus dem Lexikon *Sector-.dic*

5.2.8 Lexika der Organisationsnamen

Organisationsnamen sind Entitäten, welche ähnlich wie Branchenbezeichnungen im Kontext eines Arbeitsverhältnisses auftreten können. Wenn es darum geht, die Zugehörigkeit zu einem Betrieb, einer Firma oder einem Konzern auszudrücken, kann man auf Organisationsbezeichner nicht verzichten. Da häufig Personennamen im Wortlaut von Firmennamen vorkommen können, ist es von Vorteil ein Organisationsnamenlexikon aufzubauen. Somit kann später ausgeschlossen werden, dass z.B. bei *„John Miller Systems“* *Systems* als zweiter Nachname erkannt wird. Außer einem Eintrag im Organistionsnamenlexikon (ONL) gibt es noch weitere Faktoren, welche sicherstellen können, dass Organisationen auch als solche identifiziert werden. So kann beispielsweise die Nennung der Rechtsform nach dem Firmennamen oder eine kurze Beschreibung des Unternehmens beim Finden eines Organisationsnamen hilfreich sein.

Allgemeine Organisationsbezeichnungen

Wie bereits Friederike Mallchok im Zuge ihrer Dissertation [Mallchok, 2004] herausgefunden hatte, werden Organisationsnamen in Wirtschaftstexten meist durch eine kurze Beschreibung eingeführt. Sie fertigte für diesen Zweck ein Lexikon für Organisationsbezeichnungen an, welches insgesamt 23.800 Einträge enthält.

Abbildung 5.10 gibt einen Einblick in dieses Organisationsbeschreibungslexikon, welches kurz ODL von ihr genannt wurde und ab jetzt im System unter dem Namen *orgbez-.dic* verwendet wird.

```
accessories retail company,.orgbez
accountancy and consultancy firm,.orgbez
automobile seat manufacturer,.orgbez
bankholding company,.orgbez
bath-solution maker,.orgbez
bicycle company,.orgbez
```

Abbildung 5.10: Auszug aus dem Lexikon *orgbez-.dic*

Eigennamen von Organisationen

Außerdem hat Friederike Mallchok ein Lexikon für Organisationsnamen (ONL) mit fast 286.500 Unternehmen angelegt, welches hier ebenfalls berücksichtigt werden soll und den Namen *org-.dic* trägt.

```
Intel Corp,.org
Kids Unlimited Ltd,.org
Caterpillar Group,.org
Deutsche Telekom,.org
Walt Disney Pictures,.org
```

Abbildung 5.11: Auszug aus dem Lexikon *org-.dic*

Mithilfe diverser Listen aus dem Internet konnte das Organisationsnamenlexikon von Friederike Mallchok um einige Einträge erweitert werden. Auch die in Abschnitt 5.2.3 beschriebene Methode der Eigennamenerkennung, welche auf die Jahresausgabe der Financial Times 2004 angewendet wurde, lieferte weitere potentielle Firmennamen. Viele der damals aussortierten Begriffe waren Bezeichnungen für Unternehmen, welche an dieser Stelle von Nutzen sein können.

Firmen, die ihrem Namen die entsprechende Rechtsform anfügen, wurden allerdings doppelt in das neue Wörterbuch *Companies-.dic* aufgenommen – einmal mit dem Kürzel für die Rechtsform und einmal ohne dieses. Dadurch kann das jeweilige Unternehmen in beiden Varianten im Text auftreten und wird jedes Mal gefunden.

Um die neu gewonnenen Lexikoneinträge in das System aufnehmen zu können, wurde ein gesondertes Wörterbuch mit mehr als 229.600 Einträge angelegt, bei dem die Einträge aus dem ONL von Friederike Mallchok nicht integriert wurden. Aus Organisationsgründen sollten die Inhalte beider Lexika nicht miteinander vermischt werden, was aber dem System keine Einbußen bringt, da ihm insgesamt 515.500 Organisationsnamen zur Verfügung stehen.

Wie Abbildung 5.12 deutlich macht, wurden den Einträgen im Lexikon *Companies-.dic* grammatikalische sowohl als auch semantische Zusatzinformationen mitgegeben.

```
Abit AG,.N+XN+PR+Company
Abitex Resources Inc,.N+XN+PR+Company
Above Chase Nominees Ltd,.N+XN+PR+Company
Above Chase Nominees,.N+XN+PR+Company
A&C Black plc,.N+XN+PR+Company
Adler GmbH,.N+XN+PR+Company
Adolf Wurth & Co KG,.N+XN+PR+Company
Advanta Corp,.N+XN+PR+Company
```

Abbildung 5.12: Auszug aus dem Lexikon *Companies-.dic*

Organisationsspezifische Adjektive

Es gibt gewisse Adjektive oder adjektivisch verwendete Partizipkonstruktionen, welche vornehmlich im linken Kontext von Organisationsbezeichnungen oder -namen auftreten. Die gebräuchlichsten organisationsspezifischen Adjektive extrahierte Friederike Mallchok aus den Kontexten der Organisationsnamen, die ihr NER-System [Mallchok, 2004] im Reuters Korpus erkannte, und fasste sie zu einem Lexikon mit 110 Einträgen zusammen. Dieses hat den Namen *org_adj-.dic* und wird auszugsweise in Abbildung 5.13 dargestellt.

```
foreign,.org_adj
newly formed,.org_adj
privatized,.org_adj
profitable,.org_adj
worldwide,.org_adj
```

Abbildung 5.13: Auszug aus dem Lexikon *org_adj-.dic*

Organisationsspezifische Kontexte

Außerdem sammelte Friederike Mallchok noch die linken und rechten prädikativen Kontexte von Organisationsnamen. Da im Zuge meiner Untersuchungen der Kontexte von Menschenbezeichnern nur Organisationsnamen in Objektposition interessant sind, werde ich lediglich auf ihr Wörterbuch *context_before-.dic* zurückgreifen, welche den linken Kontext von Firmennamen spezifiziert.

```
authorized the,.context_before
available from,.context_before
award from the,.context_before
awarded to,.context_before
```

Abbildung 5.14: Auszug aus dem Lexikon *context_before-.dic*

5.2.9 Lexika der geographischen Begriffe

Toponyme sind aus Biographien eigentlich kaum wegzudenken. Mindestens einmal wird in jedem Lebenslauf eine Ortsbestimmung genannt – der Geburtsort. Doch manchmal sind Biographien wahre Bewegungsprofile, was nichts anderes heißt, als dass die betreffende Person vielleicht oft den Ausbildungsort, den Wohnort oder den Arbeitsort gewechselt hat. Um die Erkennung von geographischen Begriffen in biographischen Kontexten zu erleichtern, ist es von Vorteil ein Toponymlexikon zu erstellen.

Aus diversen Internetquellen [MapPlanet GmbH, 2006] [Wikipedia, 2005/2006] wurden geographische Entitäten wie Länder, Kontinente, Bezirke, Provinzen, Regionen, Départements, Grafschaften, US Bundesstaaten, Bundesländer, Städte und Stadtteile extrahiert und in den beiden Wörterbüchern *GeosMapplanet-.dic* und *GeosWikipedia-.dic* gespeichert. Des Weiteren war Sebastian Nagel so freundlich, mir sein Toponymlexikon (*GeosSebastian+.dic*) zu überlassen, so dass dem System nun insgesamt 286900 geographische Entitäten zur Identifikation von Ortsbestimmungen im Kontext von Menschenbezeichnern zur Verfügung standen.

Die folgende Aufstellung 5.7 gibt einen detaillierten Überblick zu allen in den Lexika verwendeten Abkürzungen. Diese Übersicht beschränkt sich ganz auf die grammatikalischen und semantischen Zusatzinformationen für nominale Ortsbestimmungen. Toponyme, welche als Adjektive im Text fungieren, werden an späterer Stelle ausführlich behandelt.

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
N	grammatikalisch	Noun	Nomen
XN	grammatikalisch	Extended Noun	Mehrwortlexem
City	semantisch	City	Metropole, Stadt
NYCBourough	semantisch	New York City Bourough	Stadtteil von New York City
Borough	semantisch	Borough	Stadtteil, Stadtbezirk
CaProvince	semantisch	Canadian Province	Kanadische Provinz
Département	semantisch	Département	Département
County	semantisch	County	Grafschaft
Region	semantisch	Region	Region, Gebiet
USstate	semantisch	US State	US Bundesstaat
Nation	semantisch	Nation	Land
Continent	semantisch	Continent	Kontinent
GEO	semantisch	Geographical Term	Toponym, geographischer Begriff

Tabelle 5.7: Abkürzungen, die in *Geos*-.dic* für Nomina verwendet werden.

Kontinente und Länderbezeichungen

Eine relativ kleine Teilmenge der Toponyme sind die Länderbezeichnungen, welche im Lexikon *GeosWikipedia-.dic* kodiert sind und ungefähr 2,4% der insgesamt 18.000 Wörterbucheinträge ausmachen.

```
Africa,.N+Continent+GEO
America,.N+Continent+GEO
Belgium,.N+Nation+GEO
Bosnia and Herzegovina,.N+XN+Nation+GEO
```

Abbildung 5.15: Auszug aus dem Lexikon *GeosWikipedia-.dic*

Städtenamen

Das Lexikon *GeosMapplanet-.dic* [MapPlanet GmbH, 2006] besteht ausschließlich aus Städtenamen und enthält ungefähr 25.400 Einträge. Dazu kommen noch ungefähr 15.100 Städtebezeichnungen aus *GeosWikipedia-.dic* [Wikipedia, 2005/2006], welche die größte Subkategorie (nahezu 84%) in diesem Wörterbuch bilden.

```
Athens,.N+City+GEO
Baden-Baden,.N+City+GEO
Bahçe,.N+City+GEO
Cap-Haïtien,.N+City+GEO
Castellammare del Golfo,.N+City+GEO
```

Abbildung 5.16: Auszug aus dem Lexikon *GeosMapplanet-.dic*

Grafschaften, Regionen, Bezirke und Départements

Eine weitere Gruppe bilden die regionalen Bezeichnungen. Denn etwa 12% der Lexikoneinträge sind Bezeichnungen für englische Grafschaften, europäische Regionen, kanadische Provinzen, deutsche Bundesländer und französische Départements. Da Ortsnamen häufig durch ihren jeweiligen Verwaltungsbezirk spezifiziert werden, können diese gesammelten Einträge durchaus von Nutzen sein.

```
Andalusia,.N+Region+GEO
Manhattan,.N+Borough+NYCBourough+GEO
Newfoundland and Labrador,.N+CaProvince+GEO
Val-de-Marne,.N+Département+County+GEO
Warwickshire,.N+County+GEO
```

Abbildung 5.17: Auszug aus dem Lexikon *GeosWikipedia-.dic*

US Bundesstaaten und ihre typischen Abkürzungen

Die US-amerikanischen Bundesstaaten sind ihrem syntaktischen Verhalten ähnlich wie die in Abschnitt 5.2.9 genannten Gebiete. Dass sie gesondert aufgeführt werden, liegt an ihrer abgekürzten Schreibweise, die meist häufiger als die ausgeschriebene Form in Texten vorkommt. Somit ist es sinnvoll, beide Varianten – die offizielle Bezeichnung des Bundesstaates und die gebräuchliche Abkürzung – im Lexikon *USstates-.dic* aufzuführen.

```
Alaska,.N+USstate+GEO
Calif\.,.N+USstate+GEO
California,.A+AGEO+AState
Idaho,.N+USstate+GEO
Ida\.,.N+USstate+GEO
Louisiana,.N+USstate+GEO
Minnesota,.N+USstate+GEO
Minn\.,.N+USstate+GEO
```

Abbildung 5.18: Auszug aus dem Lexikon *USstates-.dic*

Geographische Adjektive

Soll der Standort eines Unternehmens näher bestimmt werden, ist es in englischsprachigen Nachrichtentexten gebräuchlicher den geographischen Begriff als Adjektiv einer Firmenbezeichnung voranzustellen. In der Regel gibt es zu jedem nominalen Toponym auch ein Adjektiv, welches in seiner grammatikalischen Form mit dem Nomen identisch sein oder von diesem abweichen kann. Auch existieren in der englischen Sprache grammatikalische Gesetze zur Bildung der Adjektivform einer Ortsbestimmung, die jedoch für die US Bundesstaaten relativ locker gesehen werden. Beispielsweise ist es grammatikalisch korrekt *„the Florida company Miller International“* in einem Text zu schreiben, doch hat sich auch die Variante *„the Floridian company Miller International“* eingebürgert. Eigentlich ist die zweite Möglichkeit grammatikalisch falsch, hat sich aber durch den ständigen Gebrauch bei Nicht-Muttersprachlern so weit verbreitet, dass man diese Form auf keinen Fall vernachlässigen sollte.

```
European,.A+AGEO
Galway,.A+AGEO+ACity
Genevese,.A+AGEO+ACity
German,.A+AGEO+ANation
Icelandic,.A+AGEO+ANation
```

Abbildung 5.19: Auszug aus dem Lexikon *GeosWikipedia-.dic*

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
A	grammatikalisch	Adjective	Adjektiv
XA	grammatikalisch	Extended Adjective	lexikalische Einheit, welche die Funktion eines Adjektivs erfüllt
ACity	semantisch	City Adjective	Adjektiv für eine Metropole, Stadt
ABourough	semantisch	Bourough Adjective	Adjektiv für einen Stadtteil
AProvince	semantisch	Province Adjective	Adjektiv für eine Provinz
AState	semantisch	State Adjective	Adjektiv für einen Bundesstaat
ANation	semantisch	Nation Adjective	Adjektiv für ein Land
AGEO	semantisch	Geographical Term Adjective	Adjektiv für ein Toponym, einen geographischen Begriff

Tabelle 5.8: Abkürzungen, die in *Geos*-.dic* für Adjektive verwendet werden.

5.2.10 Lexika der Temporalia

Was wäre ein Lebenslauf ohne Zeitbestimmungen? Fast jede biographisch relevante Information wird mit einer Zeitangabe oder einer Zeitspanne versehen. Natürlich wäre es unsinnig alle möglichen Datumsangaben in einem Lexikon zu sammeln. Die Aufgabe der Datumserkennung und des Auffindens von Zeitangaben aller Art wird später eine lokale Grammatik übernehmen, welche sich aber auf Begriffe stützt, die einen gewissen Zeitraum ausdrücken.

Monatsnamen und -abkürzungen

Für die Lokalisierung von Datumsangaben sind Monatsnamen unverzichtbar. Innerhalb eines Datums kann ein Monat in ausgeschriebener oder abgekürzter Form vorliegen, wenn das Zahlenformat außen vor gelassen wird. Des Weiteren kann im Englischen eine Monatsabkürzung mit nachstehendem Punkt oder ohne diesen gebildet werden. Deshalb ist es hilfreich ein Lexikon für Monatsnamen (*Month-.dic*) und deren Abkürzungen (*MonthAbbr-.dic*) anzulegen.

```
April,.N+Month
December,.N+Month
January,.N+Month
November,.N+Month
September,.N+Month
```

Abbildung 5.20: Auszug aus dem Lexikon *Month-.dic*

```
Apr,.MonthAbbr
Aug\.,.MonthAbbr
Dec\.,.MonthAbbr
Feb,.MonthAbbr
```

Abbildung 5.21: Auszug aus dem Lexikon *MonthAbbr-.dic*

Wochentage und ihre Abkürzungen

Analog zu den Monatsnamen wurde auch ein Lexikon für Wochentage und ihre jeweiligen Abkürzungsmöglichkeiten angelegt. Dabei wurde auf die Unterscheidung zwischen der Abkürzung und der Vollform verzichtet und lediglich `DayOfWeek` als semantisches Merkmal zur Ergänzung des Wörterbucheintrages gewählt.

```
Sat,.N+DayOfWeek
Saturday,.N+DayOfWeek
Tue,.N+DayOfWeek
Tues,.N+DayOfWeek
Tuesday,.N+DayOfWeek
```

Abbildung 5.22: Auszug aus dem Lexikon *DayOfWeek-.dic*

Weitere zeitbezogene Nomina

Das Standardlexikon, auf welches das System UNITEX [Paumier, 2004] zurückgreift, kennt leider nur sehr wenige Nomina, die einen zeitlichen Aspekt ausdrücken. Um unbestimmte Zeitangaben, wie z.B. *„in the afternoon“*, später im Korpus erkennen zu können, musste die semantische Klasse `Ntime` um einige Einträge erweitert werden.

```
afternoon,.N+Ntime
autumn,.N+Ntime
evening,.N+Ntime
fall,.N+Ntime
morning,.N+Ntime
night,.N+Ntime
today,.N+Ntime
tomorrow,.N+Ntime
winter,.N+Ntime
yesterday,.N+Ntime
```

Abbildung 5.23: Auszug aus dem Lexikon *Ntime-.dic*

5.2.11 Weitere Lexika

Sicherlich könnten noch mehr Wörterbücher erstellt werden, welche bei der Entwicklung von lokalen Grammatiken zur Erkennung von Menschenbezeichnern in biographischen Kontexten später recht hilfreich sind. Doch man sollte sich dabei auf das Wesentliche beschränken und nur Lexika für semantische Klassen (Kategorien) anlegen, welche unbedingt gebraucht werden.

Alle Wörterbücher, die bis zu diesem Zeitpunkt ausführlich beschrieben wurden, enthalten Entitäten oder Bezeichnungen, welche mit hoher Wahrscheinlichkeit im Umfeld von Menschenbezeichnern auftreten. Auf diese Weise vereinfachen sie die Lokalisierung von Menschenbezeichnern im Text und ermöglichen eine sehr genaue Beschreibung des biographischen Kontextes, in dem diese vorkommen.

Zudem sind diese Lexika aus diversen Vorüberlegungen, die mit der Analyse der Nachrichtentexte aus der Financial Times einhergingen, entstanden. Dagegen wurde das folgende Lexikon während der Entwicklung verschiedener lokaler Grammatiken angelegt.

Im Verlauf der syntaktischen Studien des Korpus zeigte sich, dass Zahlen in Wirtschaftsnachrichten häufig ausgeschrieben werden, wenn es darum geht, die Dauer eines Arbeitsverhältnisses anzugeben. Leider bietet das System UNITEX grundsätzlich nur eine semantische Kategorie für arabische Zahlen, welche intern als `NB` bekannt ist.

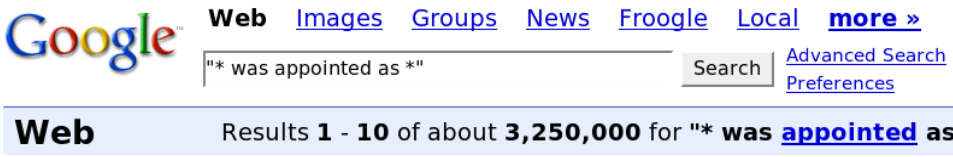
Aus dieser Not heraus entstand das Wörterbuch *NBword-.dic*, welches alle englischen Zahlen von 1 bis 100 in ausgeschriebener Form, sowie Varianten von *„hundred“*, *„thousand“*, *„million“* und *„billion“* enthält. Analog zu `NB` für arabische Ziffern wird in diesem Wörterbuch die semantische Zusatzinformation `NBword` für Zahlen in Wortform eingeführt.

```
eight,.NBword
fifteen,.NBword
twenty-two,.NBword
thirty,.NBword
forty-seven,.NBword
fifty-six,.NBword
sixty-nine,.NBword
seventy-one,.NBword
million,.NBword
one billion,.NBword
a thousand,.NBword
```

Abbildung 5.24: Auszug aus dem Lexikon *NBword-.dic*

In Anhang A auf Seite 143 ist eine komplette Übersicht aller in den Wörterbüchern verwendeten grammatikalischen und semantischen Kategorien mit ihren jeweiligen Bedeutungen und entsprechenden Erläuterungen zu finden.

5.3 Verifikationsmöglichkeiten bei Google



Google stellt wohl die größte Menge englischsprachiger Texte zur Verfügung und ist somit eine Ressource, auf die gern zurückgegriffen wird.

Was bedeutet es nun, Google als Verifikationstool einzusetzen?

Wie die obige Abbildung zeigt, ist es möglich bei Google eine Anfrage mit Asterisk zu stellen, wie z.B. `"* was appointed as *"`. Diese Schablone ermöglicht es die linken und rechten Kontexte des Prädikats *„to be appointed as“* aus Texten im Web zu ermitteln. Das führt schließlich zu Ergebnissen, wie diesen

```
In January 1990, Professor Chen San-ching was appointed as deputy director.
... and in 1997, he was appointed as Director.
Mr Thorn was appointed as Director, Office of Crime Prevention in August 2003.
```

Eigentlich bestätigen diese Ergebnisse nur die anfänglichen Vermutungen. Denn mit dieser Anfrage möchte man bestätigt bekommen, dass das Prädikat *„to be appointed as“* an seiner ersten Argumentposition eine Person in Form eines Eigennamens oder eines Personalpronomens hat, und dass auf die Präposition *„as“* eine Berufsbezeichnung folgt.

Ähnlich wie beim Bootstrapping mit lokalen Grammatiken könnte man auch hier neue Lexikoneinträge gewinnen. Nur dieses Mal geht es nicht darum, neue Informationen aus Texten zu extrahieren, sondern die Qualität einer Grammatik auf neuen Texten zu erproben. Wenn beispielsweise eine Grammatik für eine Verbrelation anhand eines bestimmten Korpus entwickelt wird, werden einige syntaktische und semantische Schlussfolgerungen für die Umgebung des Verbs gemacht. Das führt dazu, dass die Grammatik mit der Zeit so gut wird, dass fast alle Vorkommen des Ausgangswortes in diesem Korpus gefunden werden. Dabei könnten wichtige Einschübe oder Variationen in der Satzstellung übersehen werden, die in anderen Artikeln durchaus gängig sind.

An dieser Stelle kommt nun Google ins Spiel, indem die „Aussage“ der jeweiligen lokalen Grammatik konkretisiert und mit Wildcards an verschiedenen Positionen in der Phrase versehen wird. Unter der Konkretisierung einer lokalen Grammatik versteht man hier, dass ein Ausdruck der Form `:date <LN> <HumVP> as <JD>` bei Google als Anfrage nicht sinnvoll ist, dagegen aber `"* was appointed as *"` zugelassen und damit auch bestätigt wird, dass eine Instanz von `<LN>` oder `<JD>` im Umfeld einer Instanz von `<HumVP>` vorkommt.

Auf diese Weise lässt sich die Qualität der lokalen Grammatik an den entsprechenden Trefferquoten messen und die entsprechenden Passagen, welche mithilfe des Asterisks gefunden wurden, können entweder neu in der Grammatik berücksichtigt oder bestätigt werden.

6 Grammatik der Menschenbezeichner

Unter Verwendung der in Abschnitt 5.2 vorgestellten Wörterbuchressourcen lassen sich nun lokale Grammatiken zur Erkennung von Menschenbezeichnern innerhalb biographischer Relationen entwickeln.

Dabei stehen die Personen als zentrale Entität eindeutig im Vordergrund der linguistischen Untersuchungen. Aufgrund dessen werden sie am Anfang dieser strukturellen und semantischen Analyse stehen, gefolgt von weiteren Entitäten wie Organisationsnamen, Toponymen und Datumsangaben, welche im Umfeld von Personen vorkommen. Danach lassen sich die syntaktischen Zusammenhänge und biographisch relevanten Beziehungen zwischen den einzelnen Entitäten in Form von weiteren Grammatiken herstellen.

Wie bereits mehrfach angedeutet wurde, kann die semantische Kategorie der Menschenbezeichner in zwei Gruppen unterteilt werden. Diese Aufspaltung in Personennamen (Eigennamen) und in „allgemeine Menschenbezeichnungen“ wird nicht ohne Grund gemacht. Natürlich wäre eine feinere Differenzierung innerhalb dieser beiden Gruppen noch möglich, doch zeichnen sich diese zwei Subkategorien durch klassentypische, syntaktische Merkmale aus, so dass sie als zwei eigenständige Mengen betrachtet werden sollten.

6.1 Analyse von Personennamen

Eine dieser beiden Gruppen ist die der Personennamen. Leider handelt es sich hierbei um eine unendliche Menge, deren Elemente durch Kombination untereinander immer wieder neue Elemente in ihr hervorbringen. Auch können ihr jederzeit neue Namen von außen hinzugefügt werden, da im Laufe der Zeit immer mehr Bezeichnungen als neue Vornamen akzeptiert werden, welche zuvor nur bei fiktiven, literarischen Charakteren in Erscheinung getreten sind. So können beispielsweise *„Nemo“*, *„Smilla“* oder *„Aragorn“* zu gebräuchlichen Vornamen werden[40] und somit zur Expansion dieser Menge beitragen. Deshalb ist es notwendig, Regeln anzugeben, welche die Struktur von menschlichen Eigennamen beschreiben, so dass auch lexikonfremde Personennamen in Texten gefunden werden.

6.1.1 Syntaktische Variabilität bei Personennamen

Wie bereits in Abschnitt 1.3.1 auf Seite 13 angesprochen wurde, gibt es eine Vielzahl an syntaktischen Möglichkeiten, mit denen sich Personennamen darstellen lassen. Diese werden im Graphen *person_name.grf* auf Seite 71 festgehalten.

[40] Nachzulesen bei der Gesellschaft für deutsche Sprache e.V. `http://www.gfds.de/namen1.html`

Es gibt enorm viele Benennungen, welche sich auf eine einzige Person beziehen können. Das folgende Beispiel illustriert im Fall von *„Bill Gates“* wie umfangreich diese Möglichkeiten sind, obwohl es längst nicht alle von ihnen erfasst.

```
Bill Gates
Bill Henry Gates
Bill Henry Gates III
Bill H. Gates III
Bill H. Gates
William Henry Gates III
William Gates
William Henry Gates
William H. Gates III
William H. Gates
W. H. Gates
B. Gates
Mr. Gates
Mr  Gates
Gates
...
```

Abkürzung vs. Vollform

Im Beispiel von *„Bill Gates“* ist offensichtlich, dass eindeutig *„William Henry Gates III“* der Geburtsname ist, obwohl *„Bill Gates“* die gebräuchlichere Form des Namens ist. Außerdem wird deutlich, dass die syntaktische Kombination aus Kurzformen (z.B. *„Bill“*), Abkürzungen wie *„W. H.“* und Vollformen (z.B. *„William Henry“*) der Vornamen mit dem entsprechenden Nachnamen sehr viele Variationen für einen Personennamen erzeugen. Auch der Einsatz von Anredeformen und Titelbezeichnungen in Verbindung mit dem Nachnamen oder dem Vor- und Zunamen, der wiederum in einer verkürzten Form oder in der Vollform vorliegen kann, erhöht die Anzahl der syntaktischen Möglichkeiten für den Namen *„Bill Gates“*.

Synekdoche (Pars Pro Toto)

Wenn noch die Tatsache berücksichtigt wird, dass der Nachname bzw. der Vorname stellvertretend für den kompletten Personennamen im Text auftreten kann, wäre in einem formellen Dokument noch *„Gates“* bzw. *„Gates III“* und in einem persönlichen Schreiben *„Bill“*, *„William“* oder *„William Henry“* zu finden. Dieses Stilmittel ist in der Literatur unter zwei Bezeichnungen bekannt, wobei „Pars Pro Toto“ der selbst erklärende Name ist, da ein Teil des Namens für den ganzen steht. Dennoch ist es auch unter dem Namen Synekdoche geläufig.

Wenn semantische Paraphrasierungen von Namen wie z.B. *„the founder of Microsoft“* zunächst außer Acht gelassen, und Formen mit der Anrede *„Mr.“* nicht gesondert gezählt werden, so sind insgesamt über 100 Varianten des Namens *„Bill Gates“* möglich. (siehe dazu Anhang B auf Seite 145)

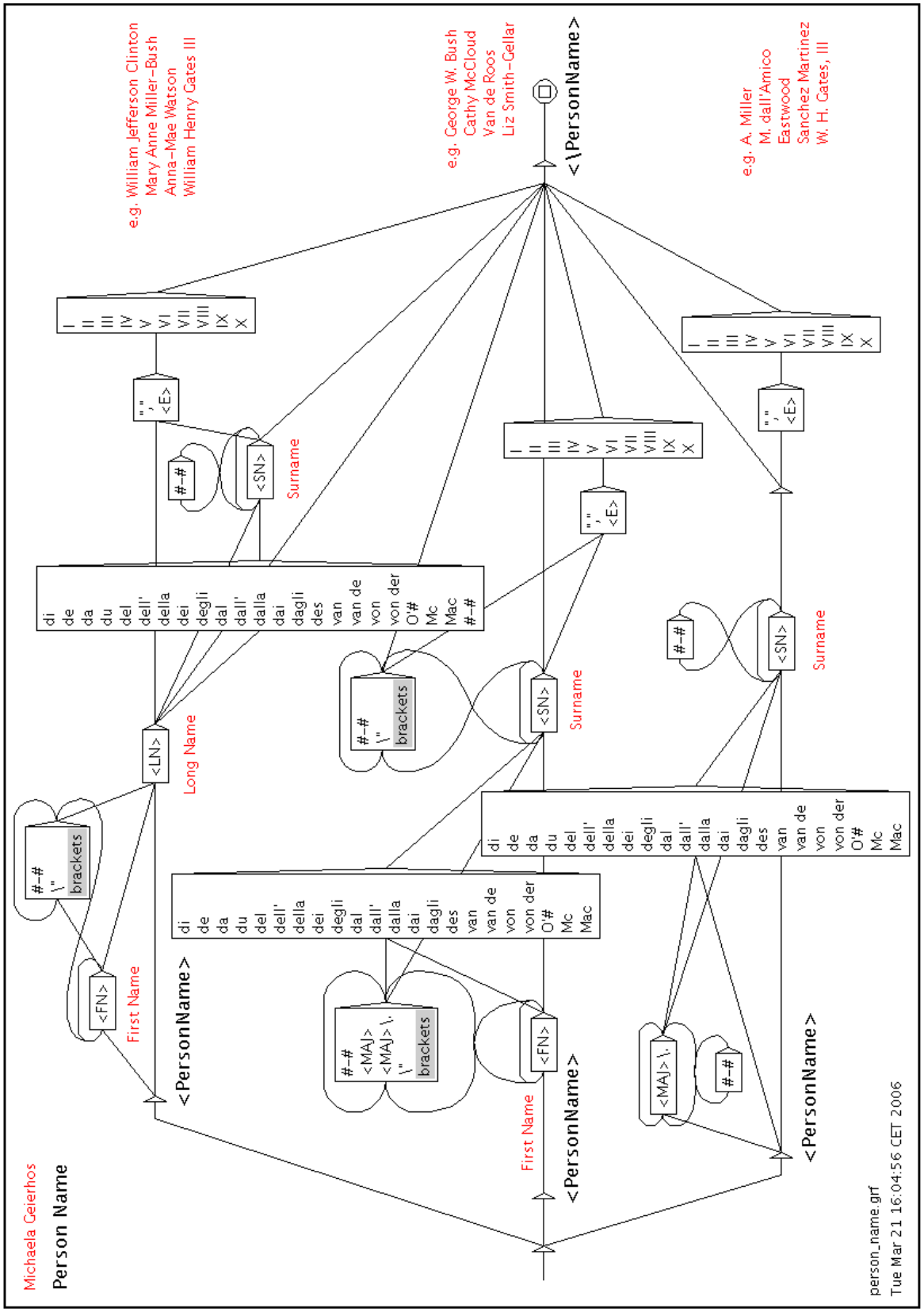


Abbildung 6.1: Graph zur Erkennung von Personennamen – *person_name.grf*

6.1.2 Disambiguierung von „Scheinnamen“

Wird der Graph aus Abbildung 6.1 ohne weitere Kontextinformationen auf das Korpus angewendet, so besteht die Möglichkeit, dass Sequenzen als Personennamen identifiziert werden, welche sich in diesem Zusammenhang nicht direkt auf eine Person beziehen.

Ein klassisches Beispiel hierfür ist die Erkennung von Personennamen innerhalb von Firmenbezeichnungen.

```
at <PersonName>Gartner</PersonName>, Inc.
```

Zwar wurde *„Gartner“* als Nachname im Text gefunden, doch fehlt die Information, dass es sich an dieser Stelle um eine Firma handelt, welche nach ihrem Gründer *„Gideon Gartner“* benannt ist.

Um diese Passage im Korpus der richtigen Entität – einer Organisation – zuzuweisen, muss der nachfolgende, sowie der vorangehende Kontext in die syntaktischen Untersuchungen miteinbezogen werden. Deshalb fungiert der Transduktor *person_name.grf* als Subgraph in der Grammatik zur Lokalisierung von Organisationsnamen, wie in Abbildung 7.1 auf Seite 80 zu erkennen ist. Dieser endliche Automat mit dem Namen *company.grf* behebt damit den vermeintlichen Fehler der Personennamengrammatik und die gesamte Sequenz wird dementsprechend im Text annotiert.

```
at <ORG><PersonName>Gartner</PersonName>, Inc.</ORG>
```

Außerdem ist auch eine Verwechslung zwischen Personennamen und Toponymen möglich, welche nur über den Kontext aufgelöst werden kann, was manchmal keine leichte Aufgabe ist, wie das folgende Beispiel illustriert.

```
<PersonName>England</PersonName>'s Queen
```

Hier wurde England fälschlicherweise als Nachname erkannt, obwohl es diese Funktion in anderen Fällen auch einnehmen kann:

```
<PersonName>England</PersonName>'s mother, Terrie, told
Secretary <PersonName>England</PersonName> served as the
Mr. <PersonName>England</PersonName> served as executive vice president
```

Wie diese Konkordanz zeigt, sollte eine nähere Spezifikation des Umfeldes von potentiellen Personennamen, die Fehlerquote deutlich eingrenzen, indem ein nachfolgender Menschenzeichner wie *„mother“*, eine voranstehende Berufsbezeichnung wie *„Secretary“* oder eine Anredeform wie *„Mr.“* sichere Indizien für die „Echtheit“ eines Personennamens sind.

Doch wird sich das Problem mit *„England's Queen“* nicht ausschließlich über den direkten Kontext lösen lassen, denn *„queen“* ist ein Menschenbezeichner wie *„mother“* und nimmt in dieser Genetivkonstruktion die gleiche Rolle ein. Nun könnte man damit argumentieren, dass immer nur ein Land an der ersten Position dieser Sequenz stehen wird, doch wäre *„Johnny's Queen“*[41] schon ein Gegenbeispiel für diese Annahme.

An dieser Stelle muss entweder der Kontext sehr detailliert beschrieben, oder es sollten Prioritäten gesetzt werden, bei denen an der genannten Vermutung festgehalten wird.

[41] *„Johnny's Queen“* ist der Name eines finnischen Rennpferdes (`http://hippos.ip-finland.com/hippos/tulokset/l20020829.php`)

6.1.3 Vervollständigung des Personennamenlexikons

Der Graph auf Seite 71 verlässt sich beim Auffinden von Personennamen ausschließlich auf Vor- und Nachnamen, sowie auf vollständige Personennamen aus den entsprechenden Wörterbüchern. Dabei wurde gezielt darauf verzichtet, Variablen für mögliche Nachnamen oder Vornamen einzuführen, weil das Risiko zu hoch ist, andere Eigennamen zu erkennen, die nichts mit Personen zu tun haben. Nur unter Berücksichtigung des Kontextes kann dieses Verfahren dazu genutzt werden, neue Personennamen zu finden, welche noch nicht in den Lexika enthalten sind.

Die Grammatik *pot_person_name.grf* arbeitet nach diesem Prinzip und versucht weitere Personennamen im Korpus zu lokalisieren.

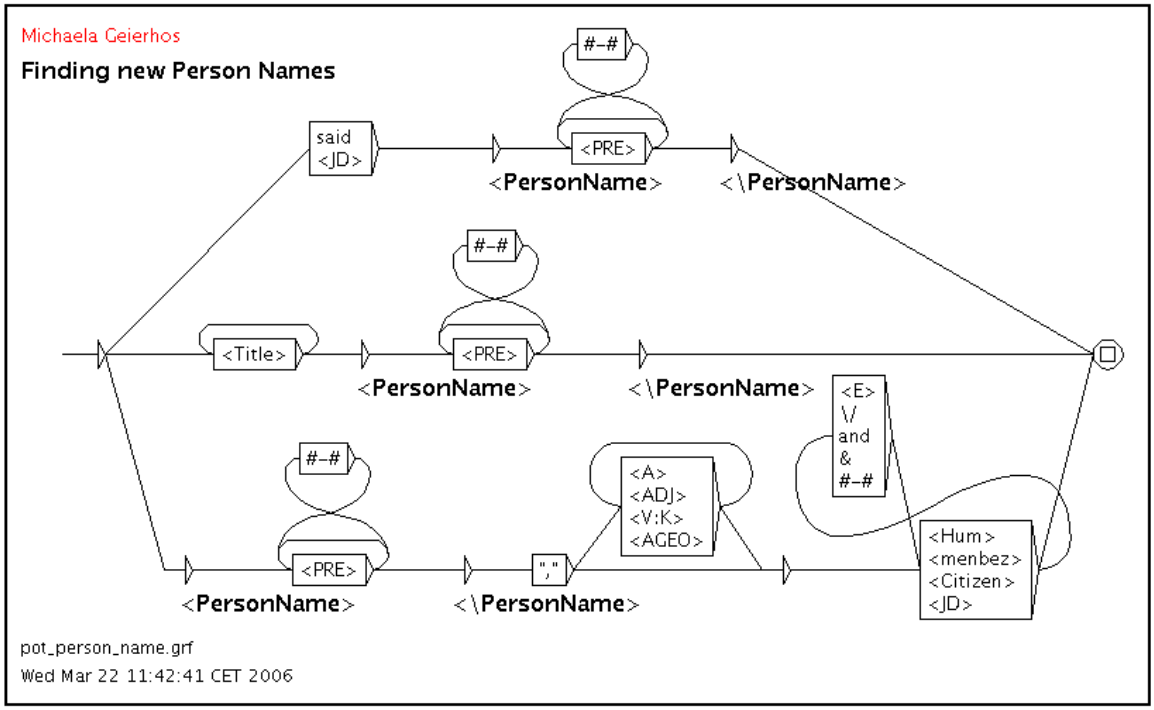


Abbildung 6.2: Graph zur Erkennung potentieller Personennamen – *pot_person_name.grf*

Bei der Suche nach möglichen Personennamen konzentriert sich dieser Graph auf die vier häufigsten Umgebungen, in denen Namen anzutreffen sind. Es werden drei Varianten in Betracht gezogen, welche den linken Kontext eines Personennamens beschreiben. Ähnlich wie schon Nathalie Friburger (siehe Abschnitt 3.3.2) Verben der Äußerung nutzte, um Personen im Korpus zu finden, wird hier stellvertretend eine Vergangenheitsform des Verbs *„to say“* verwendet.

```
said <PersonName>Allen Todd</PersonName>, Director of Utilities
said <PersonName>Clive Davis</PersonName>, Chairman and CEO
said <PersonName>James Moore</PersonName>, analyst
said <PersonName>Steve Witt</PersonName>, vice president and general manager
```

Außerdem sind Berufsbezeichnungen eindeutige Anzeichen für nachfolgende Personennamen, wenn die daran anschließenden Begriffe groß geschrieben werden.

Chief Executive Officer <PersonName>Clarence Briggs</PersonName>
chairman <PersonName>Sir Francis Mackay</PersonName>
coordinator <PersonName>Mike Trgovac</PersonName>
Coach <PersonName>Chaney</PersonName>
lawyer <PersonName>Romeo Alcantara</PersonName>
independent financial adviser <PersonName>Douglas Deakin Young</PersonName>

Auch verschiedene Anredeformen oder Titel weisen mit hoher Wahrscheinlichkeit auf nachstehende Personennamen in Form von Nachnamen oder Vor- und Zunamen hin.

Mr. <PersonName>Jose Maria Aznar</PersonName>
Mrs. <PersonName>Diana Clark</PersonName>
Prof. <PersonName>Bahram Jalali</PersonName>
Lord <PersonName>Andrew Lloyd Webber</PersonName>

Überdies sind auch im rechten Kontext von Personennamen häufig Berufsbezeichner zu finden, welche nur durch ein Komma vom Eigennamen getrennt sind.

<PersonName>Antonio Garza</PersonName>, Supply Chain Manager
<PersonName>Aurelian Lis</PersonName>, co-founder and chief operating officer
<PersonName>Bruce Chatterley</PersonName>, president and chief executive
reported <PersonName>David Breisacher</PersonName>, CEO/Chairman

So lassen sich relativ gute Ergebnisse erzielen, welche mithilfe ihrer Markierung aus der Konkordanz automatisch extrahiert, mit den vorhandenen Wörterbüchern abgeglichen und als neue Einträge in die entsprechenden Lexika aufgenommen werden können.

6.2 Allgemeine Menschenbezeichner

Der eben in Abbildung 6.2 vorgestellte Graph zur Erkennung potentieller Personennamen nutzte bereits Berufsbezeichnungen für die nähere Beschreibung des Kontextes von menschlichen Eigennamen. Daran sieht man, dass allgemeine Menschenbezeichner – in Form von Berufen, Nationalitäten, Verwandschaftsbezeichnungen, usw. – oft im direkten Umfeld von Personennamen zu suchen sind. Natürlich werden beide auch getrennt voneinander im Text auftreten. Trotz ihrer Eigenständigkeit besteht dennoch eine gewisse Verbundenheit, wodurch sich syntaktische Zusammenhänge zwischen den beiden Kategorien herstellen lassen. Unter Berücksichtigung dieser Tatsachen wurde eine gemeinsame Grammatik zur Spezifizierung von Personenbezeichnungen entwickelt.

Der Transduktor in Abbildung 6.3 auf Seite 75 vereint den Graphen der Personennamen *person_name.grf* mit der Kontextinformation aus *pot_person_name.grf* und deckt weitere Möglichkeiten ab, wie allgemeine Menschenbezeichner normalerweise im Korpus vorkommen. Dabei werden diese über die Symbole `<Hum>`, `<menbez>`, `<Citizen>` und `<JD>` in den entspechenden Wörterbüchern nachgeschlagen bzw. über den Subgraphen *jd.grf* zur Erkennung komplexer Berufsbezeichnungen (siehe Abbildung 11.4 auf Seite 119) genauer spezifiziert.

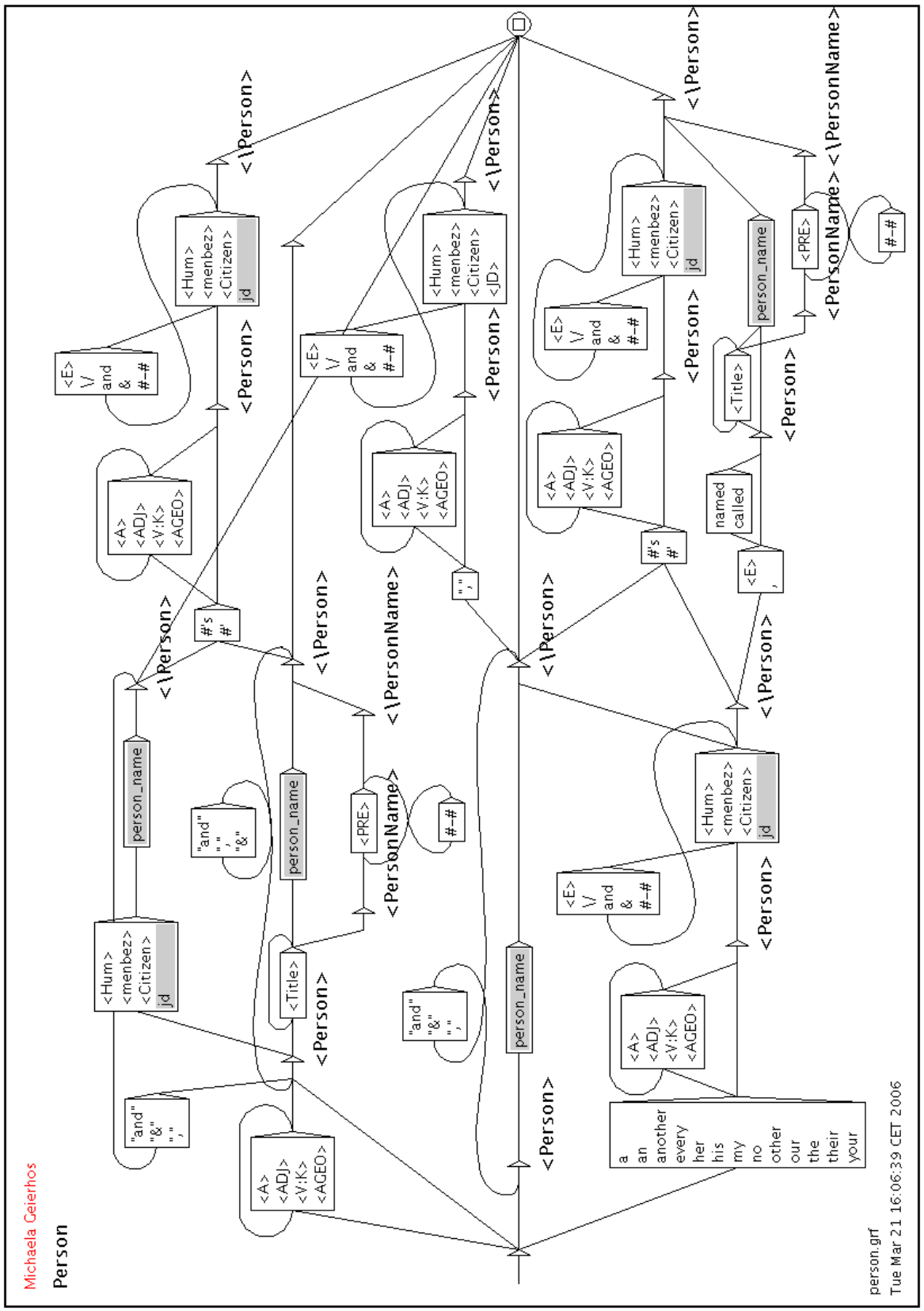


Abbildung 6.3: Graph zur Erkennung von Menschenbezeichnern – *person.grf*

said Panthers <Person><JD>defensive coordinator</JD><PersonName>Mike Trgovac</PersonName></Person>
said <Person><JD>Irish manager</JD><PersonName>Brian Kerr</PersonName></Person>
said <Person><JD>lawyer</JD><PersonName>Romeo Alcantara</PersonName></Person>, Pasig election officer
said <Person><JD>president</JD> and <JD>CEO</JD><PersonName>Barry Shaked</PersonName></Person>
<Person><JD>Spanish Prime Minister</JD><PersonName>Jose Maria) Aznar</PersonName></Person> has fostered the relationship
<Person><ORG>Senate Intelligence Committee</ORG><JD>Vice Chairman</JD><PersonName>John D. Rockefeller IV
(</PersonName></Person>
warns <Person><PersonName>Avery Shenfeld</PersonName></Person>, a CIBC World Markets economist
(She) danced with <Person><PersonName>Baryshnikov</PersonName></Person>
said <Person><PersonName>Bill Wise</PersonName></Person>, <Person>president</Person>, North American Operations
stated <Person><PersonName>Cary Losson</PersonName></Person>, <Person>Founder and President</Person> of 1031 Exchange
said <Person><PersonName>Dan Rothfeld</PersonName></Person>, <Person>senior vice president</Person>
says <Person><PersonName>Daniel Cohn-Bendit</PersonName></Person>, who led students to the barricades
stated <Person><PersonName>Eric Kuhn</PersonName></Person>, <Person>Chairman and Chief Executive Officer</Person>
said <Person><PersonName>R. Richard Fontaine</PersonName></Person>, <Person>Chairman & Chief Executive Officer</Person>
said <Person><PersonName>Steve Witt</PersonName></Person> , <Person>vice president and general manager</Person>
four years ago [Russian President] <Person><PersonName>Vladimir Putin</PersonName></Person> announced
<Person>≪JD>president</JD> and <JD>CEO</JD><PersonName>Doron Inbar</PersonName></Person> said
said <Person>lawyer<PersonName>Romeo Alcantara</PersonName></Person>
state immunity in the same way as a <Person><JD>foreign minister</JD></Person>
her face her father by pretending to be her <Person>husband</Person>
would cost the same for us, whether the <Person>traveller</Person> was booking it from Australia or France
the normally unexcitable <Person><ORG>Nokia</ORG><JD>chief executive</JD><PersonName>Jorma Ollila</PersonName></Person>

Abbildung 6.4: Konkordanz zum Graphen *person.grf*

6.3 Auflösen von Anaphern

Bis jetzt ging man davon aus, dass gezielt nach Menschenbezeichnern in Form von Eigennamen oder allgemeinen Benennungen innerhalb diverser biographischer Kontexte in Wirtschaftsnachrichten gesucht wird. Doch sollte man nicht außer Acht lassen, dass bei einer Aneinanderreihung biographischer Informationen, der Personenname wahrscheinlich nur zu Anfang und im späteren Verlauf der Ausführungen nur noch selten vorkommen wird. Stattdessen wird ein Personalpronomen die Position des Eigennamens einnehmen. Obwohl sich der Schwerpunkt dieser Arbeit auf die Erkennung expliziter Personenbezeichnungen konzentiert, muss man an dieser Stelle eingestehen, dass biographische Relationen, welche ein Personalpronomen als Subjekt haben, nicht von den hier präsentierten Grammatiken erfasst werden. Dieses Defizit ließe sich durch den Transduktor aus Abbildung 6.5 beheben, da dieser – ähnlich wie in der Diskursanalyse – versucht, die Anaphern aufzulösen, und somit die Rückbezüge der Personalpronomen auf vorangegangene Personennamen wiederherzustellen. Dabei könnte das Korpus durch den Automaten so manipuliert werden, dass das jeweilige Personalpronomen durch den entsprechenden Eigennamen ersetzt wird. Bei der Konkordanz in Abbildung 6.6 auf Seite 78 wurde allerdings noch ein Zwischenschritt vorgenommen, welcher die Zuordnung zwischen einem Personalpronomen und dem jeweiligen Bezugsnamen illustriert.

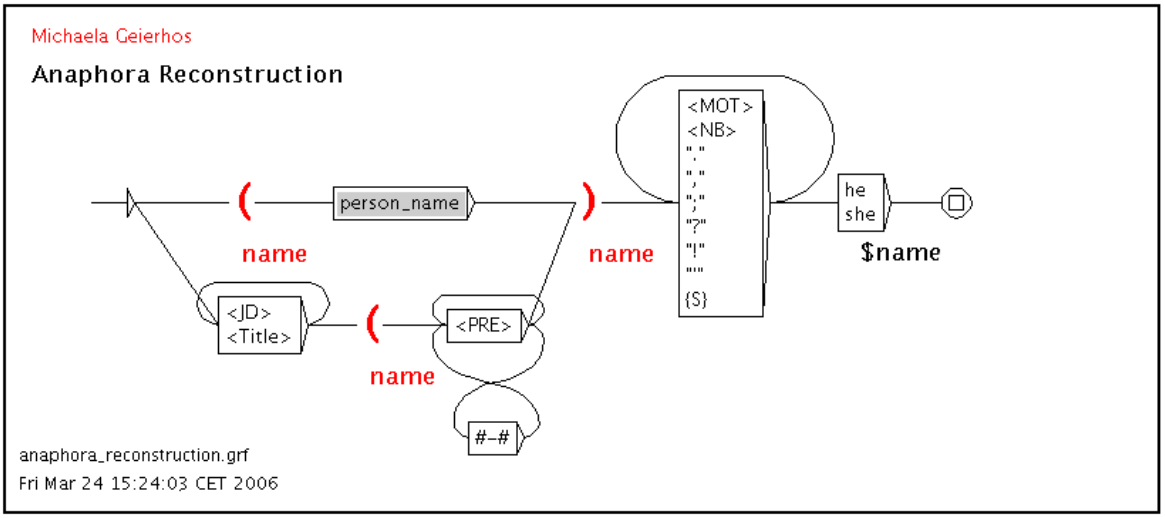


Abbildung 6.5: Graph zum Auflösen von Anaphern – *anaphora.grf*

Dennoch ist das Auflösen von Anaphern ein sehr risikobehaftetes Unterfangen, falls in den Lexika nicht das Geschlecht zu den jeweiligen Namen vermerkt wurde. Ohne diese Information lässt sich bei einem Namen nur feststellen, ob er zu einem Mann oder einer Frau gehört, wenn zuvor die Anrede *„Mr.“*, *„Mrs.“* oder *„Miss“* gefallen ist. So besteht die Gefahr, dass ein falscher Bezug zwischen dem Pronomen und dem Namen hergestellt wird. Dabei könnte beispielsweise ein weiblicher Personenname mit dem Personalpronomen *„he“* in Verbindung gebracht wird, da die Kongruenz der Geschlechter aufgrund fehlender Informationen nicht gewährleistet werden kann, und immer der letztgenannte Eigenname aufgegriffen wird.

said <PersonName>Christine Aguilera</PersonName> as (Christine Aguilera) she accepted the best female pop vocal performance
cooperation of <PersonName>Alfred Allington</PersonName> who invited us in like (Alfred Allington) he did with everyone else
I see no reason why <PersonName>Charles Allen</PersonName> should not succeed, but if (Charles Allen) he does not, then sadly that is a fact of life
I have been talking to <PersonName>Barney Allen</PersonName>, the Meath secretary, and (Barney Allen) he confirmed to me that Gary was getting an invitation
I think <PersonName>Andy Blignaut</PersonName> has been as quick of some of the others, (Andy Blignaut) he is up in the 140s
I am sure his coach <PersonName>Malcolm Arnold</PersonName> would have worked that out and I am sure (Malcolm Arnold) he has done a hard regime of quantity and quality
His son <PersonName>Aidan Barclay</PersonName> . . . gave helpful testimony but admitted at many points that (Aidan Barclay) he did not know what his father had actually done or why.
<PersonName>Craig Bellamy</PersonName> has improved every day since last Wednesday, and we will be disappointed if (Craig Bellamy) he is not there on Saturday
I think <PersonName>Berlusconi</PersonName> should take note that (Berlusconi) he is in power but not governing either the country or his own parliamentary majority If Mr <PersonName>Black</PersonName> thinks that by bringing baseless litigation (Black) he can intimidate Richard Breeden and deter him from fulfilling his duties as special counsel
If <PersonName>Tony Blair</PersonName> thought (Tony Blair) he had scars on his back from trying to reform the public services two or three years ago
Every year -- every single year <PersonName>George Bush</PersonName> has promised to create jobs and every year (George Bush) he is ended up losing them
How can we trust President <PersonName>Bush</PersonName> to create 2.6 million jobs when (Bush) he has the worst record since Herbert Hoover
If <PersonName>Francisco Carrasquero</PersonName> has a military complex, then (Francisco Carrasquero) he cannot be chairman of the CNE

Abbildung 6.6: Konkordanz zum Graphen *anaphora_reconstruction.grf*

7 Grammatik der Organisationsnamen

In biographischen Texten nehmen Beschäftigungsverhältnisse oft einen relativ hohen Stellenwert ein. So ist die Spezifikation eines Arbeitsverhältnisses unumgänglich, wenn biographisch relevante Informationen aus Nachrichten extrahiert werden sollen. Zunächst kann die Beziehung zwischen einer Person und einer Firma auf diese beiden Entitäten reduziert werden. Wie diese Relation nun genau aussieht, soll erst zu einem späteren Zeitpunkt geklärt werden. Deshalb muss an dieser Stelle nur die syntaktische Struktur von Firmennamen analysiert und die darin verborgene semantische Klassifizierung von Organisationen vorgenommen werden.

Natürlich werden auch die Wörterbucheinträge aus Abschnitt 5.2.8 (siehe Seite 58) in die Entwicklung einer solchen Grammatik für Organisationsnamen einbezogen.

7.1 Syntaktische Variabilität bei Organisationsnamen

Ähnlich wie schon bei den Personennamen tritt eine gewisse syntaktische Variabilität in der Struktur von Organisationsnamen auf. Allerdings kann davon ausgegangen werden, dass die Vielfalt an Variationen bei Firmennamen deutlich geringer ist, als es bei den Personennamen der Fall war (siehe Abschnitt 6.1.1).

Der Graph aus Abbildung 7.1 auf Seite 80 behandelt diverse Möglichkeiten, in welcher Form eine Organisationsbezeichnung im Text vorkommen kann.

Einerseits können Unternehmen mit einem Zusatz im Namen genannt werden, der Aufschluss über ihre jeweilige Rechtsform oder die Art des Gewerbes gibt.

```
<ORG>Novartis AG</ORG> declined to comment on newspaper reports
Kim Manley, Chief Marketing Officer, <ORG>Allied Domecq PLC.</ORG>.
Ashkin, Executive Vice President of <ORG>America Online, Inc.</ORG>.
vice president, partner services, for <ORG>Choice Hotels.</ORG>
company news service from the <ORG>London Stock Exchange Freeport PLC</ORG>
Mike Gausling, President and CEO of <ORG>OraSure Technologies</ORG>.
```

Andererseits tritt ein Firmenname meist ohne diese Zusatzinformation im Text auf und wird dabei nur über das entsprechende Wörterbuch identifiziert, weil der Kontext nicht zur Disambiguierung herangezogen werden kann. Der Grund hierfür liegt im linguistischen Charakter des Organisationsnamens, der häufig im gleichen Zusammenhang wie ein Personenname verwendet wird. Außerdem gilt die Großschreibung sowohl für die Namen von Unternehmen als auch für die der Menschen. Somit bestehen kaum Chancen, eine eindeutige Abgrenzung zwischen Personen und Organisationen vorzunehmen, wenn nicht auf gewisse Lexikoninformationen oder firmentypische Indikatoren zurückgegriffen wird.

<ORG>3Com</ORG>'s wireless solution offers the price/performance
<ORG>Agfa</ORG> is pleased to achieve this milestone
<ORG>AOL</ORG> and its members have been very good to J Records
<ORG>ASUSTeK</ORG> are fully committed to developing a leading position
<ORG>AT&T</ORG> is pleased with the resolution
said <ORG>Commerzbank</ORG> analyst Christoph Rieger.

Des Weiteren erkennt der folgende Graph auch Ausdrücke, welche sich auf Institutionen, Abteilungen, Zweigstellen oder Lehreinrichtungen beziehen.

Jim Saunders of the <ORG>Department of Public Safety</ORG> said
said Brian Moyne, Project Manager, <ORG>Department of Transport</ORG>
According to a 1999 report by the <ORG>Institute of Medicine</ORG>
says Sir Howard Davies, director of the <ORG>London School of Economics</ORG>
said Claes Fornell, director of the <ORG>University of Michigan</ORG>'s
National Quality Research Center

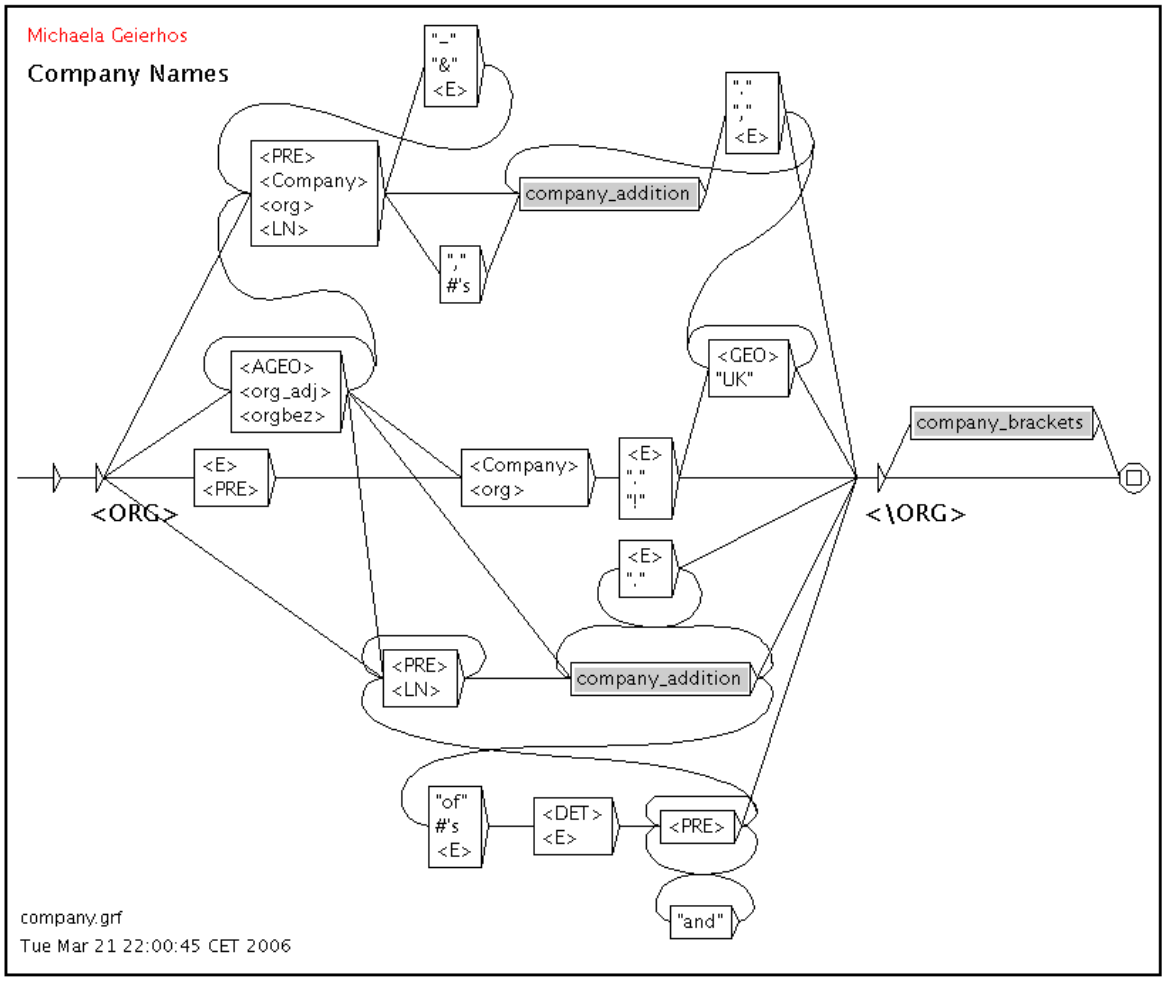


Abbildung 7.1: Graph zur Erkennung von Firmennamen – *company.grf*

Wenn das Akronym eines Firmennamens mindestens genauso bedeutend wie der eigentliche Organisationsname selbst ist, kann es durchaus vorkommen, dass es der Unternehmensbezeichnung in Klammern nachgestellt wird. In diesem Fall kommt der Subgraph *company_brackets.grf* zum Einsatz.

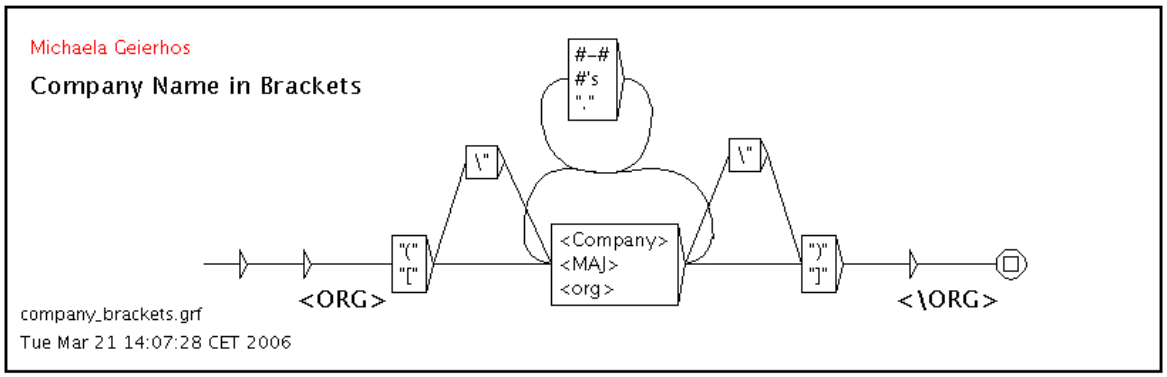


Abbildung 7.2: Graph zur Erkennung von Firmennamen in geklammerten Ausdrücken – *company_brackets.grf*

Dabei können beispielsweise folgende Textpassagen im Korpus gefunden und als Organisationsnamen gekennzeichnet werden:

```
more than 21 million <ORG>American Online (AOL)</ORG> and CompuServe members
  with production at <ORG>Maschinenfabrik-Augsburg-Nuernberg (MAN)</ORG>
```

7.2 Abgrenzung von unechten Organisationsnamen

Da sich der endliche Automat *company.grf* aus Abbildung 7.1 auf sehr eingeschränkte Kontextinformationen stützt, besteht die Möglichkeit, dass Sequenzen irrtümlich als Firmennamen erkannt werden. Demnach wären für diesen Graphen folgende Phrasen potentielle Fehlerquellen:

```
said <ORG>Gary Flood, Executive</ORG> Vice President of
     said <ORG>Glenn Jackson, co</ORG>-founder
```

Diese vermeintlichen Organisationen entsprechen dem Schema eines Personennamens gefolgt von einem Firmenzusatz. Doch war in diesem Fall keiner der nachgestellten Firmennamenindikatoren für eine eindeutige Identifikation ausreichend. Da der Subgraph *company_additions.grf* aus Abbildung 7.3 auf Seite 82 bei der Erkennung von Rechtsformen und anderen möglichen Zusätzen keine Rücksicht auf Groß- und Kleinschreibung nimmt und nicht im Text „nach vorn blickt“, werden auch nachfolgende Mehrwortlexeme zerrissen bzw. nicht erkannt. Jedoch kann dies nur passieren, wenn der Transduktor ohne zusätzliche Informationen zum textuellen Umfeld des Unternehmens eingesetzt wird. Für die hier vorgestellten linguistischen Studien von biographischen Relationen, werden Firmennamen lediglich innerhalb bestimmter Satzstrukturen untersucht. Deshalb wird die Grammatik zur Erkennung von Organisationsnamen immer nur als Subgraph – eingebettet in einem umfassenden Kontext – aufgerufen werden.

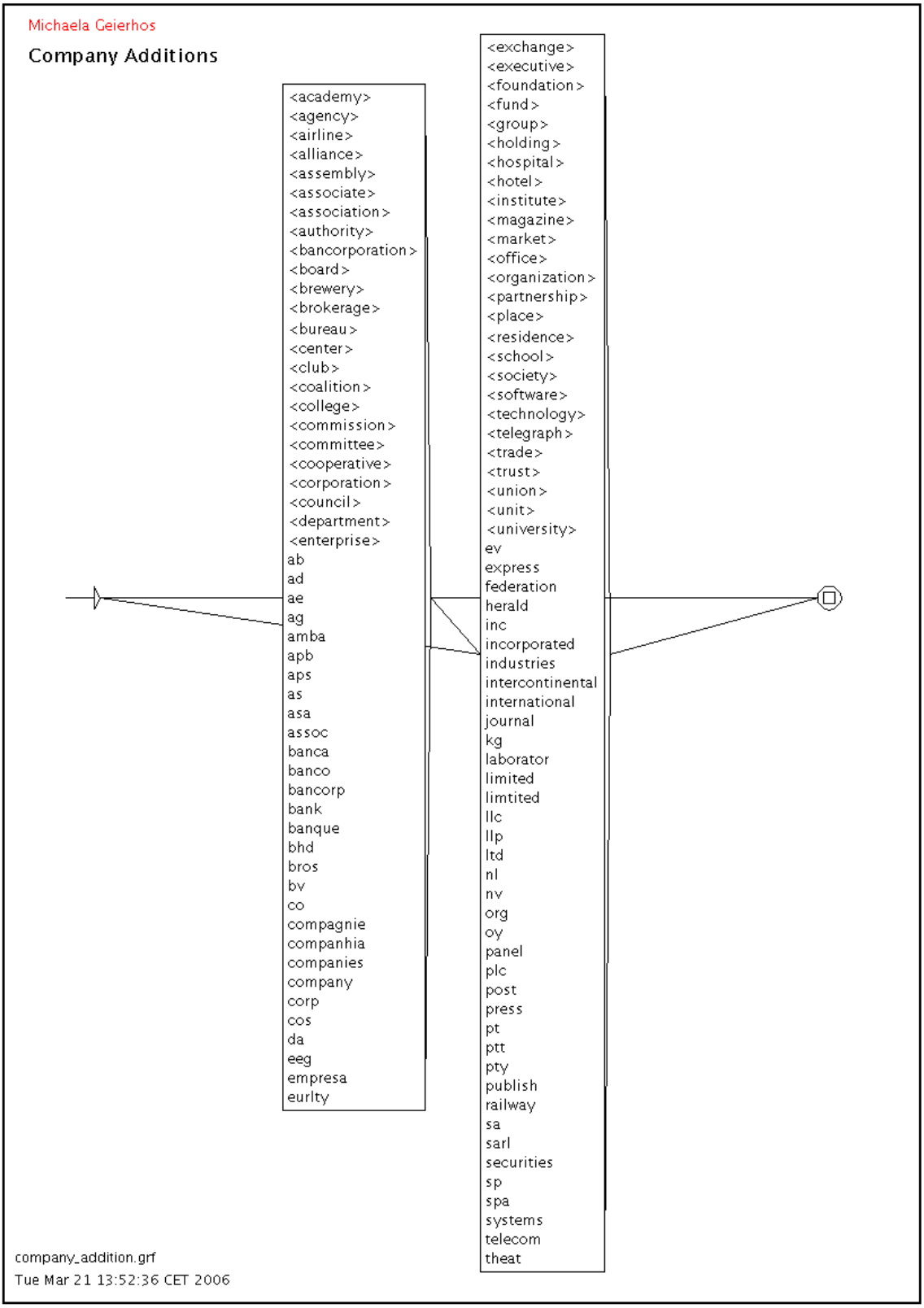


Abbildung 7.3: Graph zur Erkennung von Rechtsformen und weiteren Firmenzusätzen – *company_additions.grf*

7.3 Vervollständigung des Organisationsnamenlexikons

Ähnlich wie schon bei den Personennamen lassen sich Grammatiken auch für die Suche nach neuen Organisationsnamen einsetzen und erweitern auf diese Weise die Lexika.

Der folgende Graph bedient sich einerseits der linkstypischen Kontexte für Firmennamen, und andererseits berücksichtigt er nachstehende Firmenzusätze aus dem Subgraphen *company_additions.grf*. Dabei werden die von Friederike Mallchok [Mallchok, 2004] gesammelten Kontextinformationen in Bezug auf Organisationen verwendet, welche über die Symbole `<contextbefore>`, `<org_adj>` und `<orgbez>` abgerufen werden können.

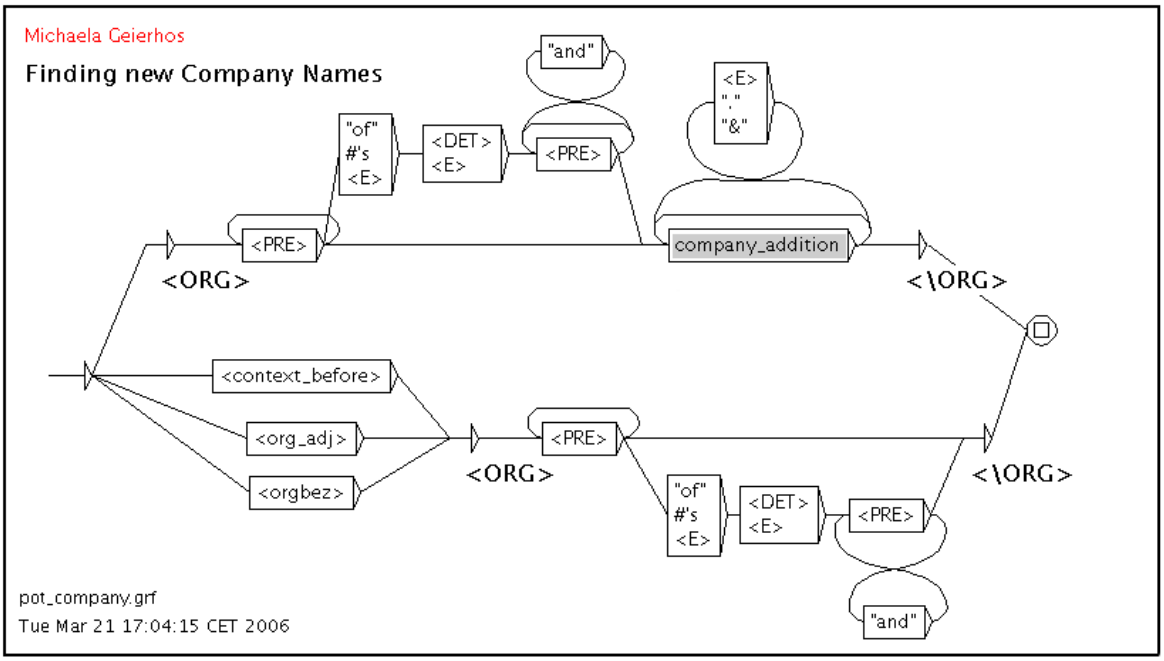


Abbildung 7.4: Graph zur Erkennung potentieller Firmennamen – *pot_company.grf*

Eigentlich ist dieser Automat schon im Graphen *company.grf* in Abbildung 7.1 auf Seite 80 enthalten. Hier ist seine Aufgabe aber, potentielle Organisationsnamen zu lokalisieren, die nicht unbedingt in einem der Wörterbücher vorkommen.

So findet er beispielsweise folgende Firmennamen im Korpus:

```
<ORG>Above All Software</ORG> was selected from the hundreds of companies
    a partner in <ORG>Accenture's Insurance Solution Group</ORG>.
  said Phil Flynn, an analyst with <ORG>Alaron Trading Corp.</ORG>
stepping in as a white knight to rescue <ORG>Aventis SA</ORG> from
      President and CEO of <ORG>OraSure Technologies</ORG>.
president of <ORG>Mobile Competency Inc.</ORG>, a consulting firm.
```

Die Treffer lassen sich nun automatisch aus der Konkordanz extrahieren, sollten aber noch sorgfältig durchgesehen werden, bevor sie mit den bestehenden Lexikoneinträgen abgeglichen werden. Als neue Wörterbucheinträge können sie später dabei helfen, die Qualität der Grammatik zu verbessern.

8 Grammatik der Ortsangaben

Geographische Begriffe treten in verschiedenen Zusammenhängen in Texten auf. Dabei ist die häufigste Aufgabe eines Toponyms die Angabe eines Ortes, an dem ein Ereignis stattgefunden hat oder stattfinden wird. In Anbetracht dessen, dass in dieser Arbeit Menschenbezeichner innerhalb biographischer Kontexte untersucht werden sollen, lassen sich die hier verwendeten Lokativa eher auf Geschehnisse in der Vergangenheit beziehen.

8.1 Biographische Relationen mit Ortsangaben

In der Menge der biographischen Relationen gibt es einige Prädikate, die meist zusammen mit einer Ortsangabe genannt werden.

Darunter fallen beispielsweise die Verben *„to be born“* und *„to be raised“*:

Artemis Wines president Eric Smith, born and raised in <GEO>Chile</GEO>, heads Artemis Wines International

Born in <GEO>Long Island, New York</GEO> on March 27, 1970, Carey moved to New York City at the age of 17

Die hier erkannten Toponyme wurden mithilfe des Automaten aus Abbildung 8.1 auf Seite 87 im FT-Korpus gefunden. Jedoch fungiert diese Grammatik nie als eigenständige Komponente bei der Lokalisierung von Ortsangaben, da sie grundsätzlich im Kontext anderer Transduktoren aufgerufen wird. Im obigen Beispiel wurde sie als Subgraph in den Graph der Verbalphrase *„to be born (and raised)“* eingebunden (siehe Abbildung 10.2 auf Seite 99).

Für das Auffinden geographischer Begriffe sind in der Grammatik einige „Wörterbuch-Lookups“ notwendig. Die entsprechenden Lexika, welche dazu herangezogen werden, wurden bereits in Abschnitt 5.2.9 auf Seite 61 eingeführt.

So wird einerseits mit folgenden semantischen Wörterbuchkategorien gearbeitet:

- `<Bourough>` – sucht nach einem Stadtteil.
- `<CaProvince>` – sucht nach einer kanadischen Provinz.
- `<City>` – sucht nach einer Stadt.
- `<Continent>` – sucht nach einem Kontinent.
- `<County>` – sucht nach einer Grafschaft.
- `<Département>` – sucht nach einem französischen Département.
- `<GEO>` – sucht nach Toponymen, die keine besondere Kennzeichnung haben.

- `<NYCBourough>` – sucht nach Stadtteilen von New York City.
- `<Nation>` – sucht nach Ländern.
- `<Region>` – sucht nach Regionen.
- `<USstate>` – sucht nach U.S. Bundesstaaten.

Andererseits werden noch folgende Schlüsselwörter zur Disambiguierung des Kontextes verwendet:

beach, bourough, city, county, country, district, province, region, state, town

In der Umgebung dieser Wörter sind mit hoher Wahrscheinlichkeit weitere geographische Begriffe zu finden, die eventuell noch nicht in einem der Lexika enthalten sind. So können im selben Schritt neue und bekannte Toponyme im Korpus gefunden werden.

Ein weiteres Indiz für eine nachfolgende Ortsangabe ist eine Himmelsrichtung wie *North*, *West*, *East*, *South* bzw. *Northern*, *Western*, *Eastern*, *Southern* oder andere ortsbegrenzende Adjektive wie *Middle* und *Central*.

Auf diese Weise lassen sich die unmittelbaren Kontexte von Toponymen auf die wesentlichen Bestandteile eingrenzen und tragen dazu bei, die Quote der falsch erkannten geographischen Begriffe zu minimieren.

8.2 Ortsangaben in ihrer Funktion als Attribute

Des Weiteren beziehen sich Lokativa nicht nur auf Ereignisse, sondern sie bestimmen auch Nominalphrasen näher, indem beispielsweise Aufgabengebiete von Menschen spezifiziert bzw. eingegrenzt werden, oder der Standort einer Firma als Zusatz in deren Namen wiedergegeben wird.

8.2.1 Toponyme als Attribut einer Berufsbezeichnung

Für die Erkennung von Berufsbezeichnungen ist zwar der Graph *jd.grf* auf Seite 119 zuständig, doch ruft dieser in seinem linken und rechten Kontext die Toponymgrammatik *geo.grf* auf. Somit werden beispielsweise folgende Berufsbezeichnungen im Umfeld von Ortsangaben lokalisiert:

```
said Peter Gregory, <GEO>England</GEO>'s chief medical officer.
said EMC product manager for <GEO>South Asia</GEO> Ajaz Munsiff
```

8.2.2 Toponyme als Attribut eines Organisationsnamens

Analog dazu sucht der endliche Automat *company.grf* auf Seite 80 im rechten Kontext von Firmennamen nach geographischen Begriffen, welche den Sitz des Unternehmens angeben.

```
or of travel and fleet services for J&J <GEO>Europe</GEO>.
  Dean Tang, president and CEO of ABBYY <GEO>USA</GEO>.
```

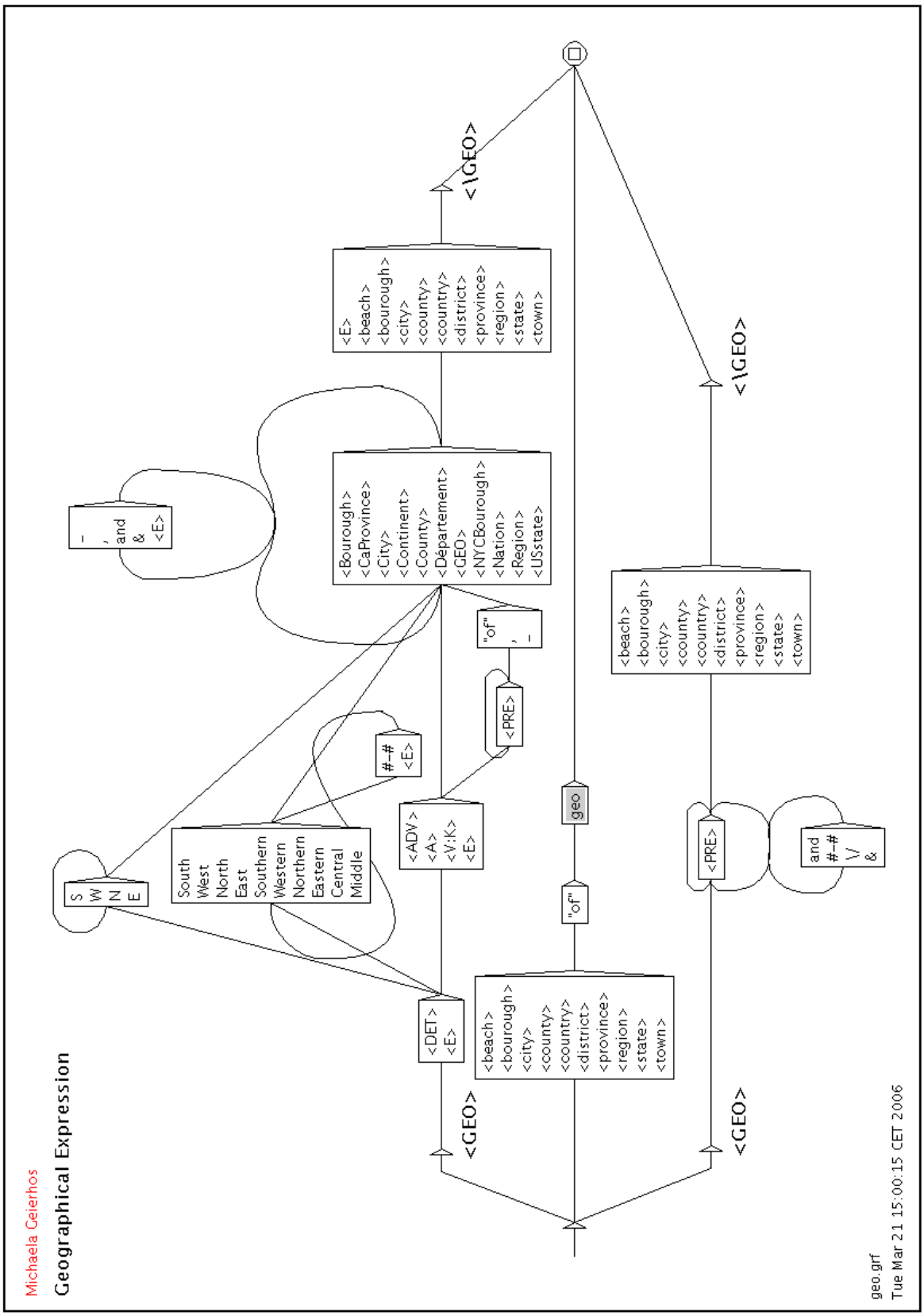


Abbildung 8.1: Graph zur Erkennung von Toponymen – *geo.grf*

AMC-10 will be delivering some of <GEO>America</GEO>'s leading cable programs
Caricom observers be in <GEO>Antigua</GEO> and <GEO>Barbuda</GEO> at this time
said EMC product manager for <GEO>South Asia</GEO> Ajaz Munsiff
the traveller was booking it from <GEO>Australia</GEO> or <GEO>France</GEO>
for example to add a few branches in <GEO>California</GEO>
said Peter Gregory, <GEO>England</GEO>'s chief medical officer.
the Tripartite Aggression against <GEO>Egypt</GEO> was launched in 1956.
officer of Prescribed Solutions, a New York</GEO>-based cosmeceuticals company.
the blowing up of homes in <GEO>Moscow</GEO> and <GEO>Volgodonsk</GEO> (which has never been proved).
Dean Tang, president and CEO of ABBYY <GEO>USA</GEO>.
Naomi Schwartz, SBCAG Chairwoman and <GEO>Santa Barbara</GEO> County Supervisor (First District).
Economic Group, a research firm in <GEO>Lansing</GEO>, <GEO>Mich.</GEO>.
companies have said they plan to list in <GEO>London</GEO> but they are a long way from it
an analyst with Alaron Trading Corp. in <GEO>Chicago</GEO>.
Dr Tien Wu, President of ASE Americas, <GEO>Europe</GEO> and <GEO>Japan</GEO>.
A mission of Delta Dental Plan of <GEO>Tennessee</GEO> is to support programs of demonstrated value
manager Brian Little, aiming to emulate <GEO>Chesterfield</GEO> and <GEO>Wycombe</GEO>, Second Division clubs
A commodities boom threatens to give <GEO>Canada</GEO> a bad case of the Dutch disease
House Democratic leader Nancy Pelosi (<GEO>Calif.</GEO>) and other Democratic congressional leaders.
a market research consultancy based in <GEO>Austin</GEO>, TX</GEO>.
Sales Vice President Diana Clark, <GEO>Los Angeles</GEO>.
that Italy's exclusion from the <GEO>Berlin</GEO> meeting reflects a loss of influence over EU affairs
number of companies in both <GEO>Germany</GEO> and <GEO>Switzerland</GEO> have expressed an interest
A car in <GEO>San Francisco</GEO> would cost the same for us

Abbildung 8.2: Konkordanz zum Graphen *geo.grf*

9 Grammatik der Datumsangaben

Bereits von Anfang an war klar, dass ein System zur automatischen Erkennung biographischer Relationen nicht ohne eine lokale Grammatik für Datumsangaben auskommen wird. In Anlehnung an die Grammatiken, welche Maurice Gross zur Beschreibung genauer und ungefährer Datumsangaben erstellt hatte [Gross, 1993], wurde ein Finite State Graph entwickelt, welcher Datumsangaben innerhalb biographischer Relationen als solche identifiziert. Da in biographischen Kontexten meist präzise Datumsangaben gemacht werden, wird nur der entsprechende Referenzgraph von Maurice Gross in Abbildung 9.1 gezeigt. Wie nun der Graph für die Erkennung von Daten innerhalb biographischer Relationen in Wirtschaftsnachrichten aussieht, illustriert Abbildung 9.2 auf der folgenden Seite.

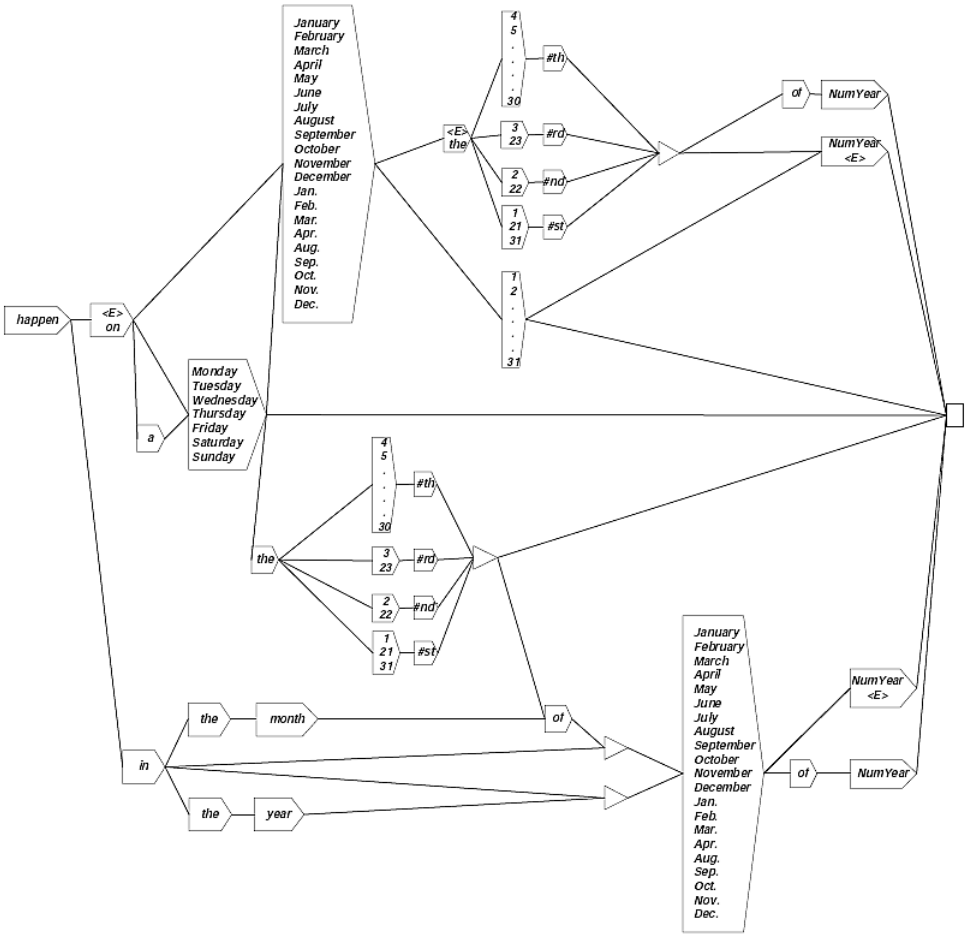


Abbildung 9.1: Graph zur Erkennung genauer Datumsangaben aus [Gross, 1993]

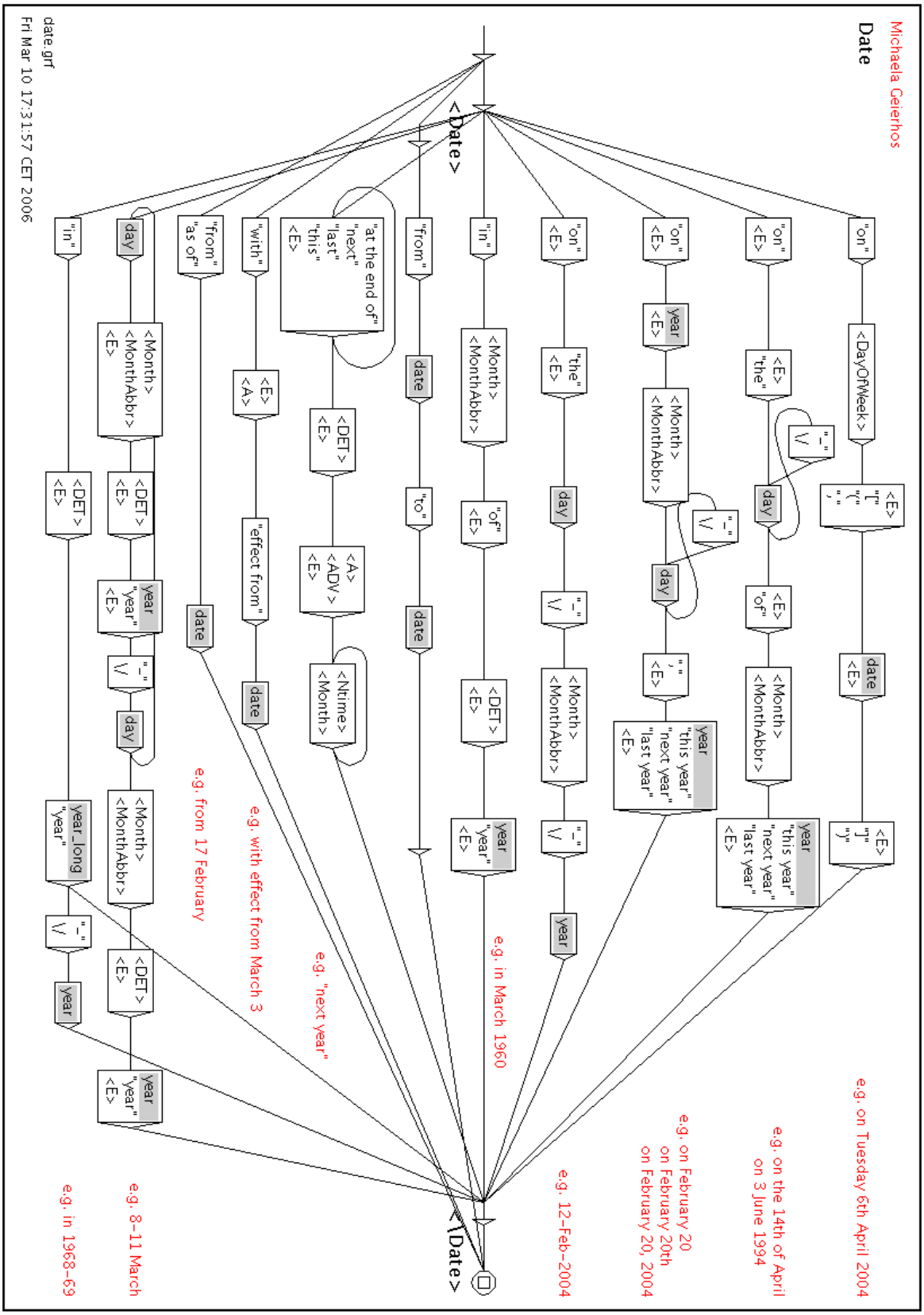


Abbildung 9.2: Graph zur Erkennung von Datumsangaben – *date.grf*

Da der Graph *date.grf* aus Abbildung 9.2 ein sehr komplexes Gebilde ist, lässt sich seine Vorgehensweise bei der Lokalisierung von Datumsangaben zunächst schwer nachvollziehen. Aufgrund dessen ist es durchaus sinnvoll die Arbeitsweise dieses Transduktors Schritt für Schritt nachzuvollziehen.

Dabei werden die wichtigsten Pfade des Graphen virtuell durchlaufen, um die Funktionsweise des Automaten zu erläutern. Betrachtet man den ersten (obersten) Pfad des Graphen *date.grf* isoliert, so lassen sich deutlich zwei Varianten von Zeitangaben erkennen, welche dieser Automat beschreibt. Einerseits werden damit Textpassagen entdeckt, die Aufschluss darüber geben, an welchem Wochentag etwas geschehen ist, und andererseits wird diese Information noch mit einem genauen Datum versehen.

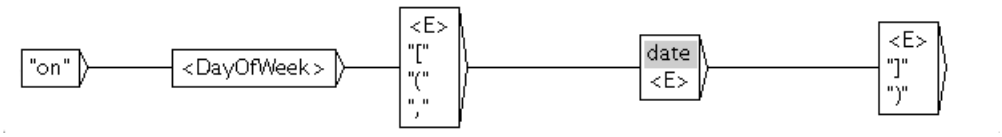


Abbildung 9.3: Erster Pfad aus dem Graphen *date.grf*

So würde dieser Ast des Graphen beispielsweise Ausdrücke der Form

```
on Saturday
on Monday, September 9, 2004
on Tuesday [Feb 4, 2005]
on Friday (January 7th, 1996)
on Wednesday May 3rd, 1998
```

matchen, wobei in diesem Pfad ein rekursiver Aufruf des gesamten Graphen *date.grf* erfolgt. Das heißt nun, dass dieser Ast andere Wege im Graphen miteinbezieht, welche wiederum die Erkennung des konkreten Datums vornehmen. Dabei ist der reguläre Ausdruck `on <DayOfWeek>` für die Spezifikation des Wochentages zuständig, so dass beim Knoten `<DayOfWeek>` im entsprechenden Lexikon nachgeschlagen wird, welche Wochentage an dieser Stelle möglich sind (siehe Abschnitt 5.2.10).

Die nachfolgenden Knoten können entweder leer durchlaufen werden (`<E>`) oder es wird im grau hinterlegenden Knoten `date` der Graph wieder selbst aufgerufen, um ein genaues Datum zu erkennen, welches möglicherweise geklammert auf den Wochentag folgt.

```
rname Beier, was due to appear in court  <Date>on Monday<\Date>  on two counts of blackmail charges, sai
lanthropist who came to their community  <Date>on Monday<\Date>  and handed out cash to everyone he met.
sitional government, Bahnam Ziya Bulus,  <Date>on Saturday [31 January]<\Date>  arrived in Iran via Khos
ry: President Mohammad Khatami's office  <Date>on Saturday [31 January]<\Date>  took out sections of the
results on record by the VLL board here  <Date>on Saturday,<\Date>  the company posted a turnover of Rs
charya.{S} Talking to presspersons here  <Date>on Saturday,<\Date>  Mr Bhattacharya said the bank has al
rls and four boys to the transit center  <Date>on Saturday,<\Date>  " said rebel spokesman Daya Master, a
```

Abbildung 9.4: Konkordanz zum ersten Pfad aus dem Graphen *date.grf*

Der zweite Pfad (von oben gezählt) im Graphen *date.grf* beschreibt nur eine Möglichkeit, wie ein konkretes Datum im Englischen aufgebaut sein kann.

```
31 May 1983
4th Sept 2004
on the 2nd of March 2005
on 5 Jan. 1998
on 19-20 June 2001
on 22/23 April last year
```

Dabei werden zwei Subgraphen miteinbezogen, welche den Tag und die Jahreszahl spezifizieren. Auf diese Weise wird sichergestellt, dass keine Zahl größer 31 irrtümlicherweise als Tagesangabe identifiziert wird, und dass nur richtige Jahreszahlen als solche erkannt werden.

Auf Seite 96 zeigen die Abbildung 9.12 und 9.13 die entsprechenden Automaten zur Erkennung von numerischen Tagesangaben und Jahreszahlen, welche hier nur über `day` und `year` in den jeweiligen Knoten referenziert werden.

Des Weiteren werden über die Ausdrücke `<Month>` und `<MonthAbbr>` alle möglichen Monatsnamen und Monatsabkürzungen in den beiden Wörterbüchern *Month-.dic* und *MonthAbbr-.dic* nachgeschlagen (siehe Abbildung 5.20 und 5.21 auf Seite 64).

Dabei sind auch wieder ϵ-Transitionen möglich (`<E>`), welche beispielsweise garantieren, dass sich ein Datum eingeleitet von *„on“*, als auch ein Datum ohne vorangehende Präposition finden lässt.

Die Konkordanz zu diesem Pfad des Transduktors illustriert, welche Form die Daten haben müssen, um von diesem Ast des Graphen entdeckt zu werden.

```
e new rates will remain in effect until  <Date>31 March 2004<\Date>  .{S} Source: RIA news agency, Moscow
vice - United Kingdom Government News 14 <Date>9 February<\Date>  The minister for trade and industry, D
News National Security Summits & Talks 8 <Date>9 February<\Date>  Peshawar: Shah Raza, a Pakistani broad
measures against the chemical beginning  <Date>9 June 2003<\Date>  , requiring the importers to pay cash
meyni landed at Tehran Mehrabad airport  <Date>on 1 February 1979<\Date>  .{S} Meanwhile, similar ceremon
ews Service Sports General News 90 site  <Date>on 1 February<\Date>  A mysterious disease has killed thr
an newspaper The Sunday Vision web site  <Date>on 1 February<\Date>  The International Criminal Court (I
```

Abbildung 9.5: Konkordanz zum zweiten Pfad aus dem Graphen *date.grf*

Der nächste Weg durch den Automaten ist sehr ähnlich zum Muster des zweiten Pfades, nur dass hier die einzelnen Komponenten des Datums untereinander vertauscht wurden. Wie schon der vorherige Zweig des Graphen die Angabe mehrerer Tage innerhalb eines Datums berücksichtigt hat, lässt auch dieser die Verbindung von zwei Tagen durch `'/'` oder `'-'` zu.

Der vierte Pfad des Graphen *date.grf* entspricht in seiner internen syntaktischen Struktur der des zweiten Astes. Jedoch wird hier auf eine andere Verknüpfungsmethode zwischen Tag, Monat und Jahr Wert gelegt. In Nachrichtentexten erscheint oft am Anfang oder am Ende des Artikels die Datumsangabe als `Tag-Monat-Jahr`, so dass auch diese Möglichkeit für ein gültiges Datum in Betracht gezogen werden muss.
So beschränkt sich dieses Suchmuster auf Daten folgender Form:

```
9-July-1956
on 24-Sept-2001
12/Mar/1995
on 23-October-2003
```

Im Financial Times Korpus wiesen die Datumsangaben zu Beginn oder Ende des Berichtes grundsätzlich dieselbe Struktur auf, wie auch der folgende Ausschnitt aus der Konkordanz für diesen Pfad des Automaten zeigt.

```
<Date>31-Jan-2004<\Date>   Talk of the Town# Why remember the CPP purges?{S} Business Services Political
<Date>31-Jan-2004<\Date>   The Nuclear Noose Around Pakistan's Neck Executive Legislative & General Gene
<Date>31-Jan-2004<\Date>   The Price Of Winning At Any Cost washingtonpost.com Elections Government News
<Date>31-Jan-2004<\Date>   They Strut, They Fret, They Build! washingtonpost.com 1821 BAGHDAD As Ayad Al
<Date>31-Jan-2004<\Date>   Travel# Lanuza through the rain Philippine Daily Inquirer Government News 120
<Date>31-Jan-2004<\Date>   Turn of the tide - Prepackaged Software Chemical Preparations NEC Plastics Ma
<Date>31-Jan-2004<\Date>   Why We Did not Get the Picture Executive Legislative & General Legislative Bo
```

Abbildung 9.6: Konkordanz zum vierten Pfad aus dem Graphen *date.grf*

Der folgende Zweig des Graphen *date.grf* beschränkt sich auf Daten, in denen nur Monat und Jahr genannt werden, und ganz auf den Tag verzichtet wird. Dabei werden diese Datumsangaben mit der Präposition *„in“* eingeleitet, worauf der Monatsname und eventuell im Anschluss eine Jahreszahl folgt.

Außerdem lässt dieses Suchmuster auch relative Jahresangaben zu, so dass nicht nur explizite Jahreszahlen wie z.B. *2006* sondern auch *„this year“*, *„last year“* usw. erlaubt sind. Diese Eigenschaft des Automaten wird einerseits durch den Knoten mit der Markierung `<DET>` und andererseits durch den folgenden Knoten, der das Wort *„year“* enthält, gewährleistet. Hierbei gleicht der Transduktor jede Instanz von `<DET>` – eines Determinativs – aus dem entsprechenden Wörterbuch mit dem Text ab.

```
me responsibility for the peace process <Date>in November<\Date> after he lost the defense ministry, b
id.{S} The current nuclear crisis began <Date>in October of 2002<\Date> after US officials said the No
ic" about holding six-party Korea talks <Date>in February<\Date> Executive Legislative & General Inter
 the International Atomic Energy Agency <Date>in December<\Date> last year confirmed that some individ
India exploded its first nuclear device <Date>in May 1974<\Date> .{S} Executive Legislative & General G
ed his formal announcement in the state <Date>in September<\Date> , then abandoned it until winning Iow
s annually.{S} WASA officials said that <Date>in October 2002<\Date> , they mailed an 11-page brochure
```

Abbildung 9.7: Konkordanz zum fünften Pfad aus dem Graphen *date.grf*

Der sechste Pfad im Graphen *date.grf* ist nicht nur für die Erkennung von Datumsangaben sondern auch für Zeitspannen zuständig. Da dieser Weg durch den Automaten zwei Aufrufe von *date.grf* selbst enthält, werden konkrete Daten durch andere Zweige des Transitionsnetzes gefunden, wenn ihnen die Schlüsselworte *„from“* und *„to“* vorangehen.

So lokalisiert dieser Ast des Graphen Textpassagen der Form:

```
agestan, takes up the story: "Overnight from <Date> 30 November<\Date> to <Date> 1 December<\Date> , t
In the latest figures, tourist arrivals from <Date> January 2003<\Date> to <Date> September 2003<\Date>
ipt of applications for IT scholarships from <Date> January 31<\Date> to <Date> March 31<\Date> . Acco
S} The Johor Arts Festival will be held from <Date> March 7<\Date> to <Date> April 4<\Date> . It will
a leading car and truck rental company, from <Date> November 1997<\Date> to <Date> February 2000<\Date>
```

Abbildung 9.8: Konkordanz zum sechsten Pfad aus dem Graphen *date.grf*

Der nächste Pfad im Automaten beschreibt verschiedene Möglichkeiten, wie Monatsangaben im Text zur Spezifikation von Zeitpunkten eingesetzt werden können. Dabei würde dieser reguläre Ausdruck zu folgenden Zeitangaben passen:

```
at the end of May
at the end of this beautiful spring
yesterday
tomorrow morning
last summer
```

Durch seine ϵ-Transitionen (siehe Abbildung 9.9 auf Seite 94) werden verschiedene Variationen ungefährer Zeitangaben zugelassen.

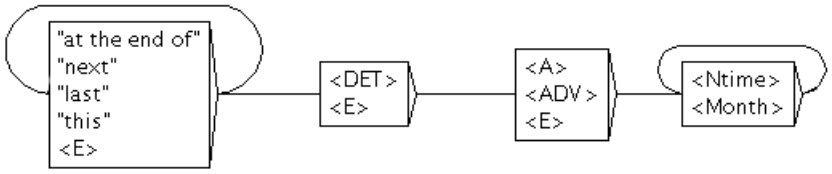


Abbildung 9.9: Siebter Pfad aus dem Graphen *date.grf*

So können Monatsangaben, sowie relative Tagesangaben wie *„tomorrow“* oder Zeitangaben wie *„morning“* und Saisonangaben wie *„spring“* für Jahreszeiten berücksichtigt werden. All diese Nomina werden über die beiden Kategorien `<Month>` und `<Ntime>` in den entsprechenden Lexika nachgeschlagen und mit dem Korpus verglichen. Welche Einträge das Lexikon *Ntime-.dic* enthält, wurde bereits in Abschnitt 5.2.10 (siehe Seite 65) erläutert.

Dagegen beschäftigt sich der achte Pfad im Graphen *date.grf* mit der Erkennung von exakten Datumsangaben, welche durch Varianten der Floskel *„with effect from“* eingeleitet werden.

Unter Variationen dieses feststehenden Ausdrucks versteht man:

```
with effect from
with immediate effect from
with restrospective effect from
with retroactive effect from
with backdated effect from
```

Ähnlich wie schon im vorherigen Ast des Graphen, ruft sich im neunten Pfad der Graph *date.grf* rekursiv selbst auf. Dieser Weg durch den Automaten unterscheidet sich vom achten Pfad nur darin, dass eine andere Floskel bzw. Präposition dem Datum vorangeht.

Der zehnte Pfad (siehe Abbildung 9.10) beschäftigt sich wieder mit der Erkennung von Zeiträumen. Ähnlich wie beim *„from ... to ...“* Zweig werden auch hier zwei Daten erkannt, die im Zusammenhang miteinander stehen. Dabei muss es sich nicht um zwei vollständige Datumsangaben handeln, denn die ϵ-Übergänge (`<E>`) lassen zu, dass beispielsweise vom ersten Datum nur der Tag genannt wird und dafür das zweite komplett im Text erscheint. So ist es möglich, Daten der Form `3/4 February 2005` , sowie `25 Oct - 10 Nov 1993` oder `2nd - 3rd May 2001` im Korpus zu finden, was die Konkordanz aus Abbildung 9.11 bestätigt.
Die nun folgenden syntaktischen Variationen zu einer bestimmten Zeitspanne, sollen anschaulich deutlich machen, wie eine ϵ-Transition `<E>` dazu beitragen kann, möglichst viele Möglichkeiten eines „Von-Bis-Zeitraumes“ zu erfassen.

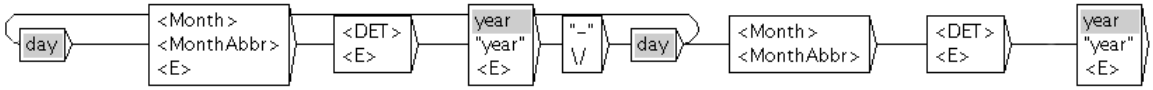


Abbildung 9.10: Zehnter Pfad aus dem Graphen *date.grf*

```
1   February      2005 - 5   February      2005
1st February      2005 - 5th February      2005
1   Feb           2005 - 5   Feb           2005
1   Feb           05   - 5   Feb             05
1                      - 5   Feb           2005
1st                    - 5th Feb.          2005
1st February last year - 5th February next year
1                      / 2   Feb           2005
1                      / 2   Feb      this year
```

Der Zyklus im Graphen, welcher durch einen Übergang vom Knoten mit der zweiten Referenz auf den Subgraphen `day` auf den ersten Knoten mit derselben Markierung entsteht, ermöglicht es, auch Daten mit folgender Struktur im Text zu lokalisieren:

```
1/2/3   Feb 2005
1/2/3/4 Feb 2005
```

Auf diese Weise können auch Datumsangaben gefunden werden, welche eigentlich eine Zeitspanne ausdrücken, in denen die einzelnen Tage separat aufgeführt sind.

```
 Ex-Date 30 January 2004 Period Covered <Date>1 January 2004 - 1 February 2004<\Date> Payment Date 2 F
ing Char Chinu District on the night of <Date>17/18 January<\Date> .{S} It is worth remembering that th
ws agency weekly schedule of events for <Date>2 - 8 Feb 04<\Date> BBC Monitoring Service - United King
ons this year will last for three days, <Date>2-4 February<\Date> , declared by [Turkmen] President Sap
al Sir Jock Stirrup, visited Ukraine on <Date>21-23 January<\Date> .{S} A broad range of issues on Ukra
ayhem.{S} The drama will be staged from <Date>22 - 26 April<\Date> .{S} For chamber music enthusiasts,
pay an official visit to Turkey between <Date>22-24 February 2004<\Date> .{S} Schroeder is scheduled to
-2004 German chancellor to visit Turkey <Date>22-24 February<\Date> Executive Legislative & General Ge
coincides with National Engineers Week ( <Date>22-28 February<\Date> ), which seeks each year to broaden
n Conference and Exhibition, Singapore, <Date>22nd-24th March<\Date> .{S} Norwegian Ship Finance Confer
4 BBC Monitoring Iranian media roundup ( <Date>25 Jan-2 Feb<\Date> ) Executive Legislative & General Leg
Feb-2004 Russian TV highlights for week <Date>26 January-1 February 2004<\Date> Television Broadcastin
tion is on a working visit to Laos from <Date>27-31 January<\Date> at the invitation of the LPRP's Dep
bruary].{S} Chen, who visited Indonesia <Date>29 January - 1 February<\Date> at the invitation of the
 a NATO high delegation to Macedonia on <Date>3-6 February<\Date> was the reason for Monday's [2 Febru
```

Abbildung 9.11: Konkordanz zum zehnten Pfad aus dem Graphen *date.grf*

Der letzte und unterste Pfad des Automaten ist für die Erkennung von Jahresangaben zuständig. Dabei werden einzelne Jahre, sowie Jahrzehnte, und auch Jahresspannen von diesem regulären Ausdruck berücksichtigt. Außerdem ruft dieser Pfad des Transduktors den Subgraphen *year_long.grf* auf, welcher im Gegensatz zu *year.grf* nur vierstellige Jahreszahlen erkennt. Analog dazu gibt es auch den Subgraphen *year_short.grf*, welcher hier nicht explizit zum Einsatz kommt, weil *year.grf* schon seine Funktion, zweistellige Jahresangaben zu finden, übernimmt. Die entsprechenden Abbildungen zu diesen Graphen befinden sich auf Seite 96.

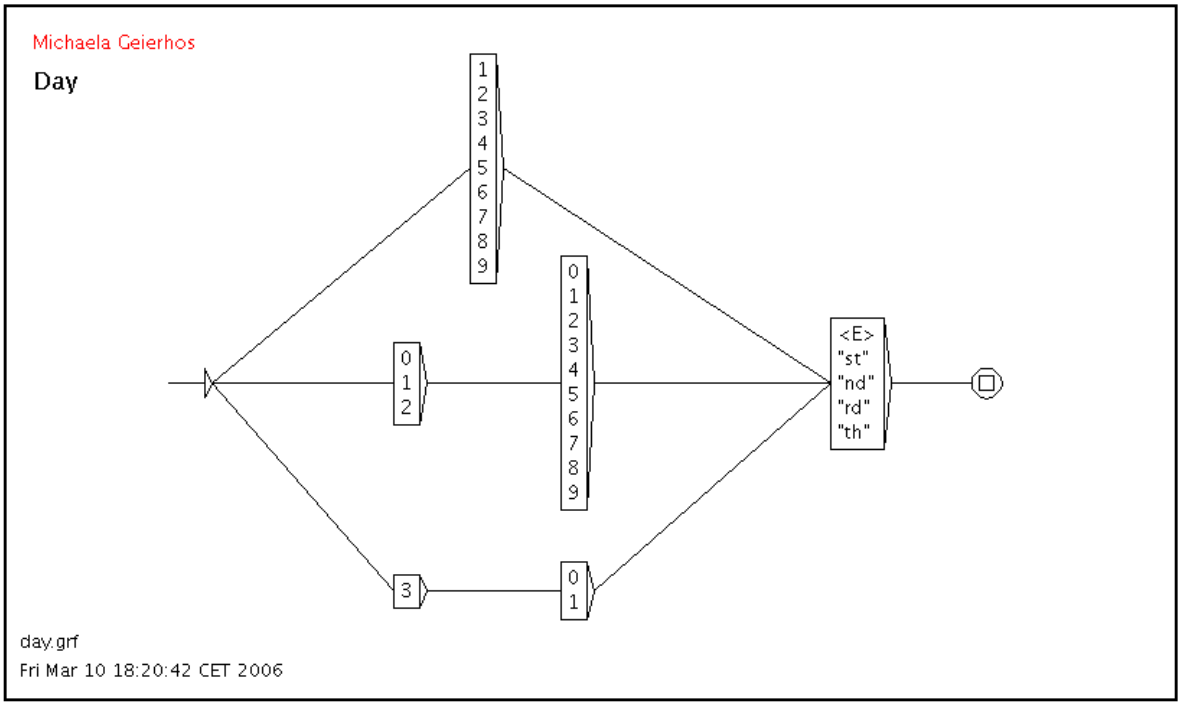


Abbildung 9.12: Graph zur Erkennung von numerischen Tagesangaben – *day.grf*

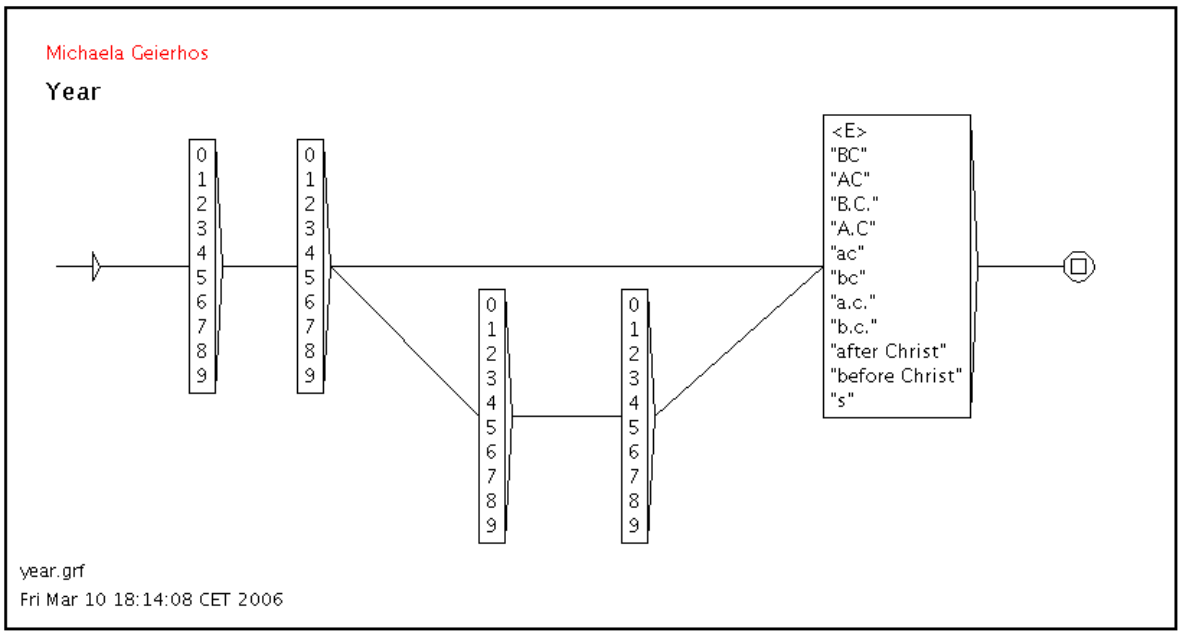


Abbildung 9.13: Graph zur Erkennung von Jahreszahlen – *year.grf*

10 Grammatik persönlicher Relationen

Nachdem in den vorangegangenen Kapiteln die Grundlagen zur Erkennung diverser Entitäten gelegt wurden, können diese nun in Beziehung zueinander gesetzt werden. Wie bereits in Abschnitt 1.2 angesprochen wurde, vermitteln die persönlichen Relationen als eine Untermenge der biographischen Relationen eine Fülle an Informationen über die verschiedensten Leute.

Da es den Rahmen dieser Arbeit übersteigen würde, auf alle englischen Prädikate einzugehen, welche in persönlichen Kontexten auftreten können, werden die wichtigsten Verben stellvertretend ausgewählt. An ihnen soll aufgezeigt werden, wie sich die unterschiedlichen Menschenbezeichner zusammen mit anderen Entitäten syntaktisch und semantisch in die Struktur dieser Relationen einfügen.

Für diesen Zweck wurde eine auf den ersten Blick sehr kompakt erscheinende Grammatik entwickelt, deren Aufgabe es ist, die einzelnen Prädikatrelationen in nicht allzu komplexen Satzgefügen zu lokalisieren.

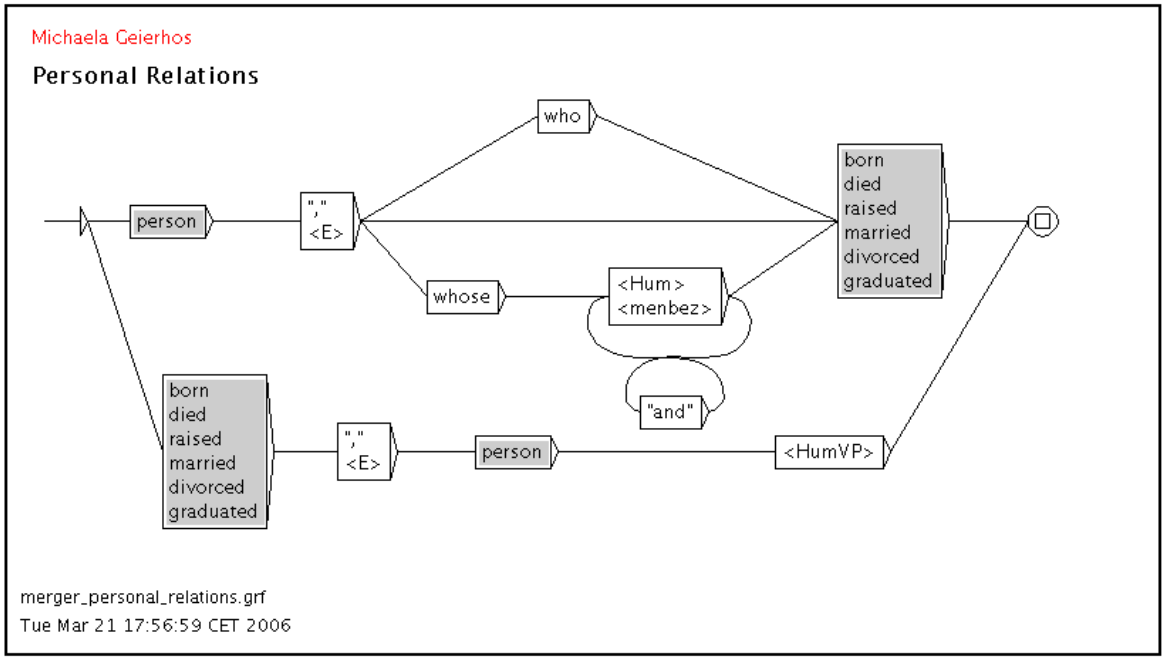


Abbildung 10.1: Graph zur Erkennung von ausgewählten persönlichen Relationen – *merger_personal_relations.grf*

Hierbei beschränkt sich der Graph aus Abbildung 10.1 auf zwei Typen von Sequenzen. Einerseits können sie mit einer Nominalphrase beginnen, in der ein Menschenbezeichner vorkommt, auf die anschließend ein Relativsatz folgt, welcher wiederum von *„who“* oder

„whose“ eingeleitet wird. Andererseits ist es auch möglich, dass die Verbalphrase, welche die persönliche Relation beschreibt, am Satzanfang steht und erst danach die betreffende Person zusammen mit einem anderen Prädikat genannt wird.

Um diese Grammatik so übersichtlich wie möglich zu halten, wurde sie sehr modular aufgebaut, so dass jeder einzelnen Relation ein eigener Graph zugewiesen wurde. Dabei decken die Subgraphen nicht nur die Verben ab, welche schon im Namen der Automaten vorkommen, sondern auch noch deren Synonyme. Auf diese Weise behandelt der Graph *merger_personal_relations.grf* 50 Verbkonstruktionen in verschiedenen Zeitformen und Variationen.

Damit erfasst dieser Automat die wichtigsten persönlichen Ereignisse im Leben eines Menschen, wozu die Geburt, die Kindheit, der Schulabschluss, die Heirat und eventuelle Scheidung, sowie der Tod gehören.

Des Weiteren besteht auch die Möglichkeit jeden der Subgraphen einzeln im Kontext des Graphen *merger_personal_relations.grf* (siehe Seite 97) aufzurufen, um sich ein Bild von der jeweiligen Relation im Satz zu verschaffen.

In den nächsten Abschnitten werden nun die verschiedenen Grammatiken vorgestellt, welche die Verbalphrasen mit den bereits genannten Prädikaten, beschreiben.

10.1 Die Geburt: „to be born“

Das Ereignis der Geburt ist wohl das am häufigsten beschriebene Geschehen in Lebensläufen. Meist werden dabei der Geburtstag, der Geburtsort und eventuell noch die Eltern der betreffenden Person erwähnt. Manchmal kann es allerdings auch vorkommen, dass jemand an dem Ort aufgewachsen ist, an dem er geboren wurde. In diesem Fall fällt der Geburtsort mit dem Ort zusammen, wo derjenige seine Kindheit verbracht hat. Somit werden auch die beiden Prädikate *„to be born“* und *„to be raised (up)“* miteinander verknüpft.

Die folgende Grammatik in Abbildung 10.2 auf Seite 99 beschreibt die syntaktischen Strukturen der Verben

- *to be born*
- *to be born and brought up*
- *to be born and raised (up)*
- *to see the light of day*

und berücksichtigt die diversen Kombinationsmöglichkeiten von Ort und Zeit im rechten Kontext der Prädikate.
Außerdem werden mit der Anwendung dieser Grammatik (siehe Seite 99) auf das Korpus einige Fragen beantwortet, welche im Zusammenhang mit der Geburt eines Menschen aufkommen.

1. Wann wurde die betreffende Person geboren?

```
Tunku Abdul Rahman was born <Date>in 1903</Date>
Cecil Beaton, who was born <Date>in 1904</Date>
their first child, Aidan Gering Pollack, born <Date>last week</Date>
```

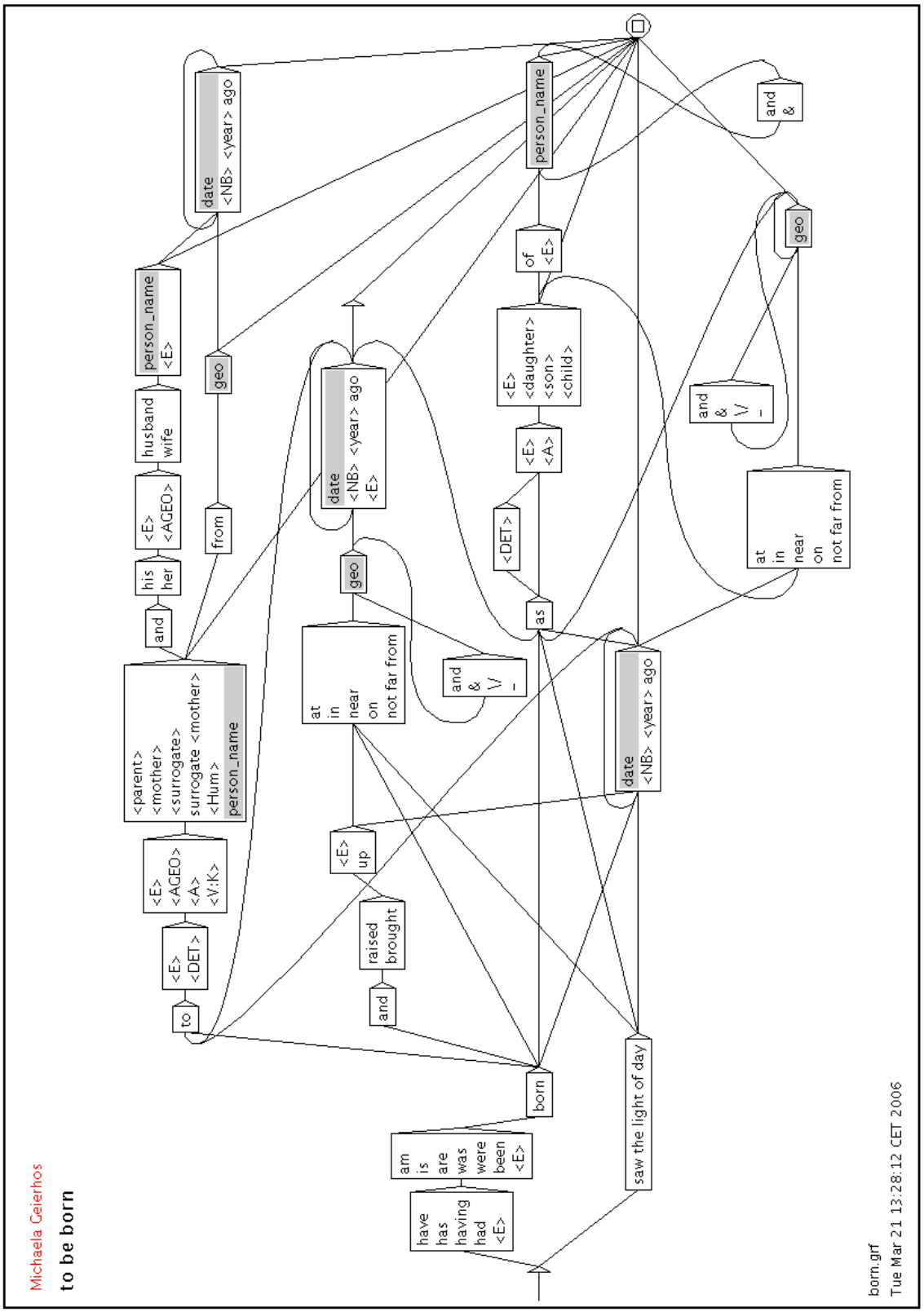


Abbildung 10.2: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to be born“* und seiner Paraphrasierung *„to see the light of day“* – *born.grf*

2. Wo wurde die betreffende Person geboren?

```
Sigman, born in <GEO>Brooklyn</GEO> in 1909
the former mayor of Tel Aviv who was born in <GEO>Germany</GEO>
McQueen, who will be 35 next month, was born in <GEO>the East End of
London</GEO>
```

3. Wer sind die Eltern der betreffenden Person (des Neugeborenen)?

```
a child was born to <PersonName>Basir</PersonName> and his Japanese wife
Prince Michael II, was born to an unknown surrogate
twin baby boys born to an American surrogate mother
Jones, who was born to Welsh parents in Papua New Guinea
She was born as daughter of Matheus Klaas and Ida Lysse in Zürich
```

4. Wie lautet der Geburtsname der betreffenden Person?

```
1934 BORN IN Poland as <PersonName>Manya Sklodowska</PersonName> Captain Jack
was born as <PersonName>Francisco Gutierrez</PersonName> in Havana/Cuba
Martika was born as <PersonName>Marta Marrero</PersonName> in Whittier, CA, on
May 18, 1969
```

Jedoch die eigentliche Frage nach dem Namen der geborenen Person wird nicht von dieser Grammatik, sondern vom Graphen *merger_born.grf* übernommen, welcher somit indirekt auch die eben genannten Fragestellungen beantwortet.

So würde nun dieser Graph Phrasen folgenden Typs beachten und darin unter anderem markieren, wer geboren wurde.

```
Born and raised in Flushing, Queens, <PersonName>Woodbridge</PersonName> had
attended the School of Visual Arts
<PersonName>Lycia Danielle Trouton</PersonName>, who was born in Belfast
<PersonName>James Saunders</PersonName> was born in London in 1925
```

Eine Konkordanz mit sämtlichen Annotationen des Transduktors *born.grf*, welcher nur Verbalphrasen mit Formen des Verbs *„to be born“* berücksichtigt, ist auf der nächsten Seite zu finden.

10.2 Die Kindheit: „to be raised (up)“

Wie bereits im vorherigen Abschnitt angesprochen wurde, ist das Aufwachsen manchmal im unmittelbaren Kontext des Geburtsortes zu suchen. Doch kann es natürlich auch als eigenständiges Geschehen auftreten und muss somit durch eine gesonderte Grammatik behandelt werden.

Der Automat, welcher nun dafür zuständig ist, herauszufinden, wo jemand seine Kindheit verbracht hat, wie derjenige aufgezogen wurde und in welchem Umfeld er aufgewachsen ist, beschäftigt sich mit einigen Verbkonstruktionen, wie sie auf Seite 102 aufgelistet sind.

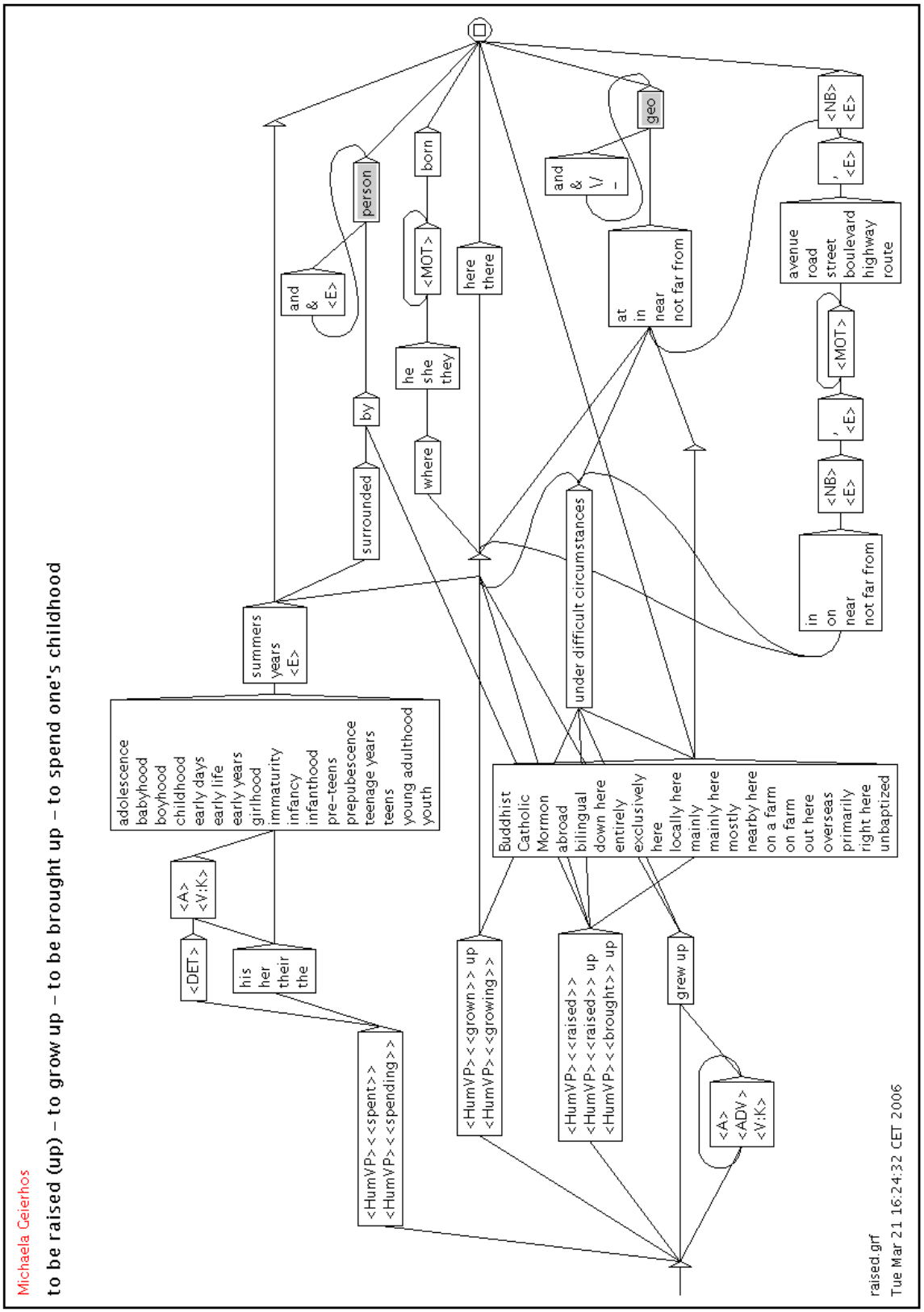


Abbildung 10.3: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to be raised (up)“* und seiner direkten Synonyme – *raised.grf*

- *to grow up*
- *to be raised (up)*
- *to be brought up*
- *to spend one's childhood*

Des Weiteren berücksichtigt dieser Graph (siehe Abbildung 10.3 auf Seite 101) die gebräuchlichsten englischen Synonyme für den Begriff der „Kindheit“. An dieser Stelle macht er sich folgendes Vokabular zunutze:

adolescence, babyhood, boyhood, childhood, early days, early life, early years, girlhood, immaturity, infancy, infanthood, pre-teens, prepubescence, teenage years, teens, young adulthood, youth

Wenn es darum geht, die Art oder den Ort der Erziehung näher zu beschreiben, stützt er sich auf folgende Ausdrücke, welche zuvor mittels Bootstrapping auf den Korpora ermittelt wurden.

Buddhist, Catholic, Mormon, abroad, bilingual, down here, entirely, exclusively, here, locally here, mainly, mainly here, mostly, nearby here, on a farm, on farm, out here, overseas, primarily, right here, unbaptized

Dagegen werden die Umstände, unter denen jemand aufgewachsen ist, meist mit der Floskel *„under difficult circumstances“* abgetan, und die Familie oder die Personen, welche in der Kindheit eine maßgebliche Rolle gespielt haben, werden häufig mit den Worten *„surrounded by“* eingeleitet.

Der Ort, an dem jemand aufgewachsen ist, kann allerdings auch sehr detailliert wiedergegeben werden, indem sogar der Straßenname genannt wird. Manchmal lässt sich auch wieder ein Bezug zum Geburtsort der betreffenden Person herstellen. Dafür wird meist die Phrase *„where so. was born“* verwendet. Zudem lässt die Grammatik noch folgende englische Synonyme für den Ausdruck “Straße“ zu:

avenue, road, street, boulevard, highway, route

Welche Verbalphrasen von diesem Graphen beispielsweise im FT-Korpus gefunden werden, illustriert die Konkordanz in Abbildung 10.4 auf Seite 103. Dagegen findet der Graph *merger_raised.grf* (ohne Abbildung) wieder heraus, wessen Kindheit und Aufwachsen hier beschrieben wird.

10.3 Der Schulabschluss: „to graduate“

Ein weiterer „Meilenstein“ im Lebenslauf kann der Schulabschluss sein. Für das Ereignis, erfolgreich von einer Schule abzugehen, werden folgende Verbkonstruktionen näher untersucht:

- *to become a graduate*
- *to receive one's degree*
- *to graduate*
- *to get one's degree*
- *to take one's degree*
- *to complete one's studies*

two young Kenyans born and brought up in <GEO>Kenya</GEO>'s coffee growing areas.
the magazine publisher James Brown, who grew up in <GEO>Leeds</GEO> and has supported the team ever since.
The 47-year old actress, who grew up near <GEO>Londonderry</GEO> and plays forensic pathologist Dr Sam Ryan
has been raised by <Person>Mrs <PersonName>Pat O'Dwyer</PersonName></Person>
have also been raised by <Person><JD>D.C. Council member</JD><PersonName>Phil Mendelson</PersonName></Person>
A woman brought up in <GEO>Australia</GEO> who became a heroine of the French resistance has finally
As a young teenager born and brought up in <GEO>Belfast</GEO>, film-maker Maeve Murphy remembers
Rosemary Canavan -- born in Scotland, brought up in <GEO>Northern Ireland</GEO>, with an English, Welsh, Irish and Huguenot background
says Matt Hollander, who grew up in <GEO> Belfast</GEO>, but lived in London for a while
Patrick Devlin grew up in <GEO>Ireland</GEO> and believes that this is the most fun band he is ever heard.
Sitwell took over Weston Hall, where he had spent his childhood, and was a much-loved member of the village.
Having spent her childhood in 48 Cherryfield Avenue, Ranelagh, just a few hundred yards from
Kevin Lewis spent his childhood being beaten by his mother at home
Nigel, who was born in 1958, spent his early years at the Mamre Brook house at the Saltram winery in Barossa
Born and raised in <GEO>Flushing, Queens</GEO>, Woodbridge had attended the School of Visual Arts
A Jewish musician raised in <GEO>Israel</GEO> and a Palestinian intellectual who
Charlotte, who was born in Derry and spent her early childhood there.
Shirley Morgan (now Shirley Larsen) was being brought up by her white <Person>parents</Person>, <Person> Jim and Jean</Person>.
I was brought up by <Person><PersonName>Bill Nicholson</PersonName></Person> at Spurs
I was brought up in <GEO> north London</GEO>.
the house where Katherine Mansfield spent her early years is closed only one day of the year.

Abbildung 10.4: Konkordanz zum Graphen *raised.grf*

Diese werden zusammen mit ihren jeweiligen rechten Kontexten im Graphen *graduated.grf* in Abbildung 10.5 detailliert beschrieben. Dabei geht diese Grammatik einerseits

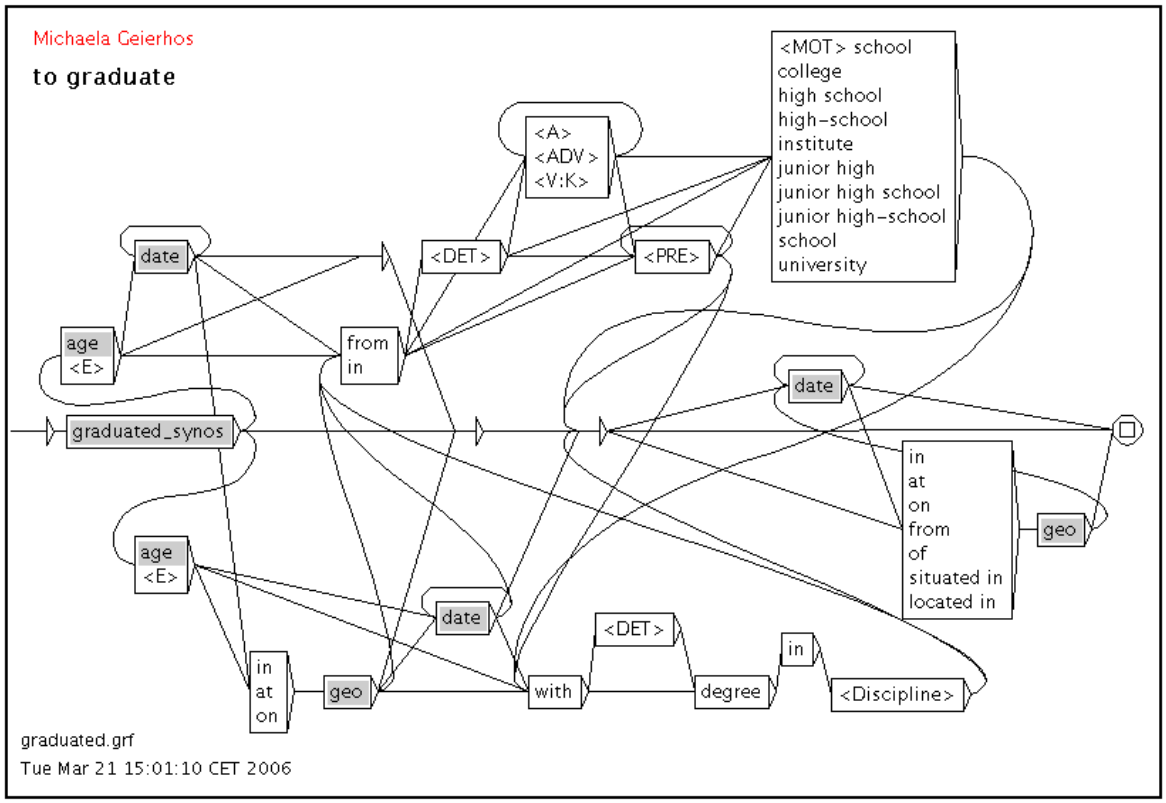


Abbildung 10.5: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to graduate“* und seiner direkten Synonyme – *graduated.grf*

auf die Schule bzw. den Namen der Lehreinrichtung ein, an welcher der jeweilige Abschluss erzielt wurde und versucht mithilfe der Schlüsselwörter

college, high school, high-school, institute, junior high, junior high school, junior high-school, school, university

vorangehende groß geschriebene Wörter als den Schulnamen zu identifizieren.

Außerdem ist es oft interessant, mit welchem Alter jemand von der Schule abgegangen ist. Um diesen Faktor in die Grammatik aus Abbildung 10.5 einzubinden, wurde ein eigener Graph zur Lokalisierung von Altersangaben entwickelt.

Dagegen werden die möglichen Verbkonstruktionen für den Ausdruck „einen Abschluss erlangen“ vom Graphen *graduated_synos.grf* beschrieben.

Doch bevor der Aufbau dieses Subgraphen genauer betrachtet wird, sollte zunächst einmal eine Beispielkonkordanz zum Hauptgraphen *graduated.grf* einen Einblick in die Art der Treffer geben.

```
                but will return home to complete his studies.
          tudents offered places in MRSM did not complete their studies.
       Andrew Gilligan had graduated from Cambridge with a degree in history.
 Conal Keaney who has also graduated from the county's All-Ireland winning under-21s
```

but those who have graduated are putting their talent and skill to good use
More than 51,000 teachers have graduated from the Enrique Jose Varona Higher Teaching Institute, which is celebrating its 40th anniversary this year. He will receive his degree <DATE>in June</DATE> in University College Cork along with Kerry footballer Mick O'Connell

An dieser Konkordanz wird deutlich, dass Informationen über die Art des Schulabschlusses, sowie die Fachrichtungen und der Name der Lehreinrichtung im direkten Umfeld von Verben mit der Bedeutung von *„to graduate“* zu finden sind.

Da die Erkennung von Schulen bereits in der Grammatik *graduated.grf* behandelt wird, werden die unmittelbar an das Verb anschließenden Phrasen, welche einen Abschluss charakterisieren, im Subgraph *graduated_synos.grf* beschrieben.

Um die Erlangung eines akademischen Grades zu spezifizieren, wurden drei Graphen entwickelt, welche die gängigsten Bezeichnungen für den Bachelor, den Master und den Doktor abdecken.

Im Gegensatz dazu wird ein Abschluss in einem bestimmten Fach über das Symbol `<Discipline>` im entsprechenden Lexikon nachgeschlagen (vgl. Abschnitt 5.2.7 auf Seite 57).

Natürlich ist das Vokabular des Automaten, welcher Bachelorabschlüsse in bestimmten Fächern erkennt, sehr ähnlich zu dem des Masterabschlussgraphen. Jedoch wurde darauf Wert gelegt, dass die darin genannten Abschlussmöglichkeiten wirklich existieren, und die in den Transduktoren enthaltenen Begriffe aus fundierten Quellen stammen [Wikipedia, 2005/2006]. Zudem ist es für eine mögliche Weiterentwicklung der Graphen ratsam, den Bachelor- und den Master-Graphen getrennt zu halten, falls neue Abschlussfächer hinzukommen sollten, in welchen man beispielsweise den akademischen Grad des Bachelors, aber nicht den des Masters erreichen kann.

10.4 Die Heirat: „to be married“

Ein weiteres Ereignis im Leben eines Menschen kann die Eheschließung sein. Aus linguistischer Sicht, ist sie für die Analyse von Personenbezeichnungen in biographischen Kontexten wesentlich interessanter als die bereits genannten Prädikate. Das liegt vorallem daran, dass eine Heirat stets zwei Personen betrifft und bei einer Aktiväußerung (X heiratet Y) beide im Satz vorkommen müssen. Zwar ist es bei dem passiven Ausdruck des „Verheiratet Seins“ nicht unbedingt notwendig, dass der Ehepartner genannt wird, doch ist dies meist der Fall.

Sowohl Aktiv- wie auch Passivkonstruktionen wurden für die Entwicklung einer Grammatik (siehe Abbildungg 10.6) bedacht, deren Ziel es ist, verheiratete Personen im Text aufzuspüren, sowie gewisse Informationen über diese Ehe herauszufiltern. Dabei wurden die folgenden Verben zum Thema „Heiraten“ oder „Verheiratet Sein“ ausgewählt:

- *to become man and wife*
- *to join in marriage*
- *to marry so.*
- *to be married to so.*
- *to get married to so.*
- *to plight one's troth to so.*

- *to pledge one's troth to so.*
- *to lead so. to the altar*
- *to take so. to wife/husband*
- *to wed so.*
- *to be wedded to so.*

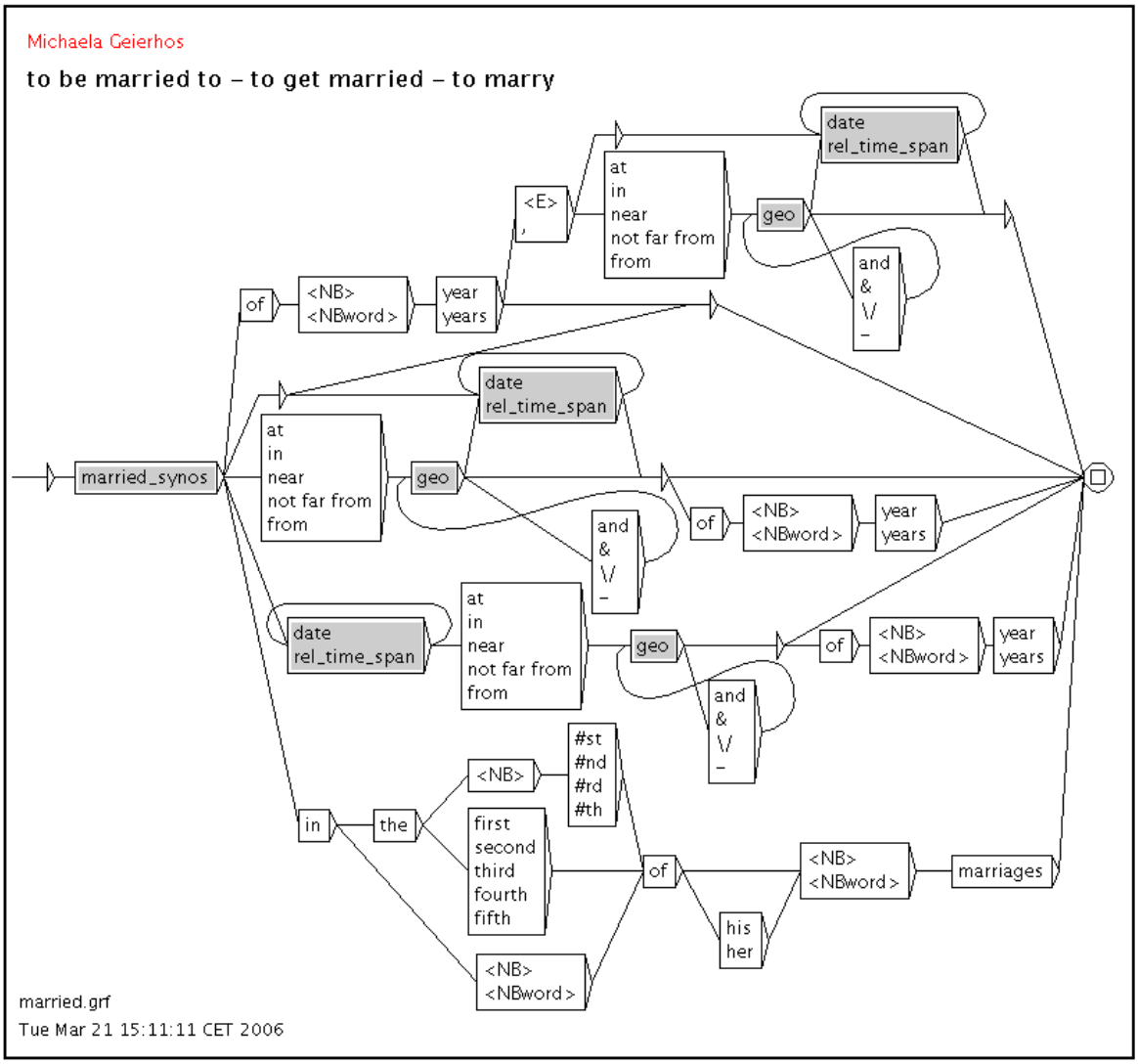


Abbildung 10.6: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to marry so.“* in seiner Aktiv- und Passivform, sowie seiner direkten Synonyme – *married.grf*

Die verschiedenen Verbformen dieser Prädikate werden zusammen mit ihren Argumenten im Graphen *married_synos.grf* in Abbildung 10.7 auf Seite 107 behandelt.

Dagegen befasst sich der Hauptgraph *married.grf* mit der Spezifizierung von folgenden Angaben bezüglich einer Ehe:

1. Wer wurde geheiratet?
2. Wann wurde die Ehe geschlossen?

3. Wo fand die Hochzeit statt?
4. Wie lange liegt die Eheschließung schon zurück?
5. Wie viele Monate oder Jahre waren sie verheiratet?
6. Um die wievielte Ehe handelt es sich?

Allerdings wird die Frage „Wer ist mit wem verheiratet?“ wieder vom Transduktor *merger_married.grf* (ohne Abbildung) geklärt.

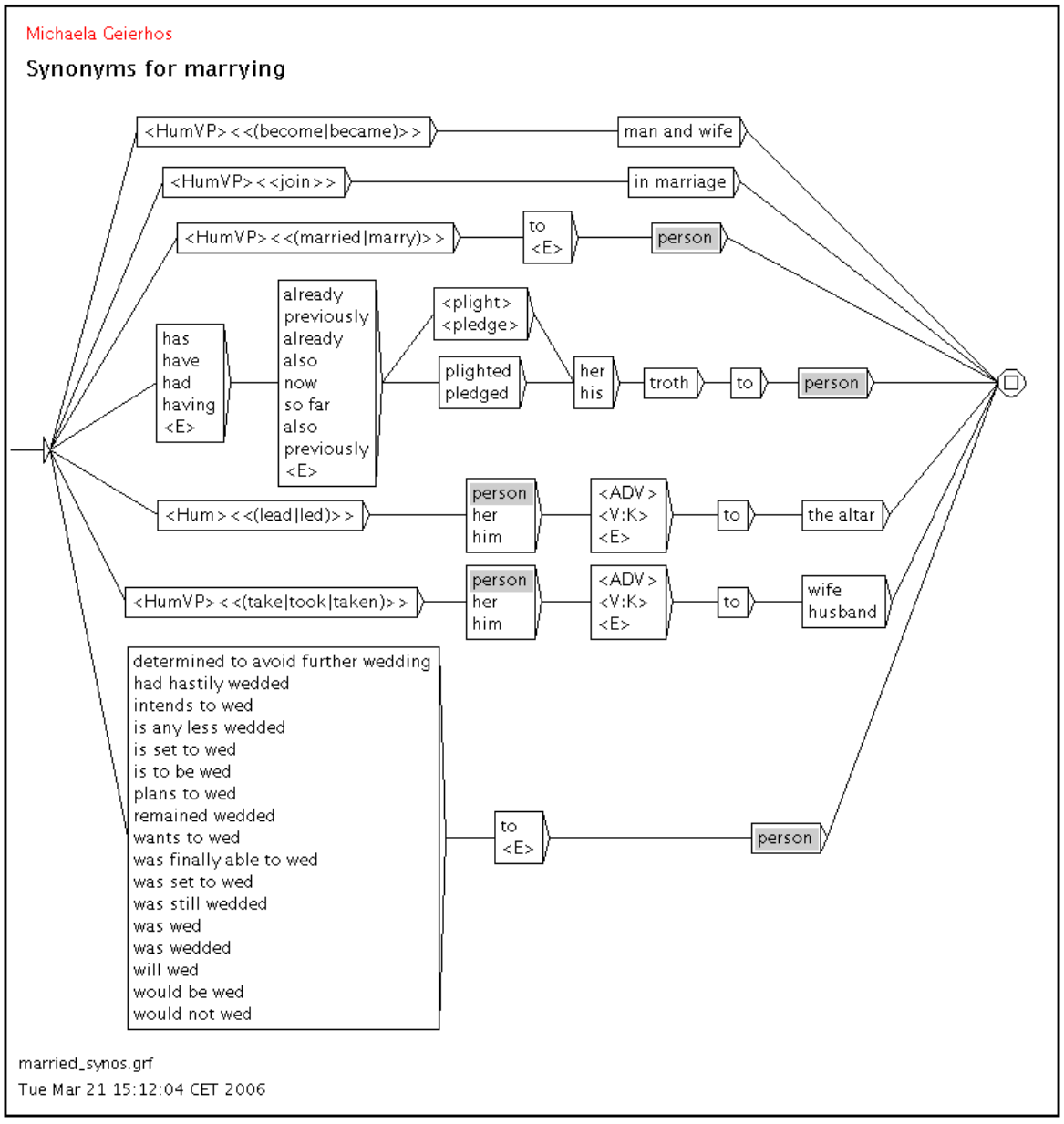


Abbildung 10.7: Graph zur Erkennung von Verben mit der Bedeutung von *„to marry, to be married“ – married_synos.grf*

Die folgende Konkordanz zeigt, welche Phrasen beispielsweise von dieser Grammatik erkannt werden.

```
                the last time a king decided to marry a <Person>divorcee</Person>
                he married <Person><JD>fashion journalist</JD><PersonName>Jackie
                                Moore</PersonName></Person>
             when Niles married <Person><PersonName>Daphne</PersonName></Person>
Agnes May immediately married a <Person><JD> wealthy businessman</JD></Person>
                Joan Collins, married to fifth <Person>husband <PersonName>Percy
                              Gibson</PersonName></Person> since 2002
                          Mary Donaldson will wed <Person>Danish Crown
                        Prince<PersonName>Frederik</PersonName></Person>
Mr Davies, who is married to <Person><PersonName>Sue Nye</PersonName></Person>
                Abraham Mendelssohn Bartholdy had married <Person><PersonName>Lea
                                Itzig</PersonName></Person>
Jason Mraz had married a <Person>friend</Person> in <GEO>Las Vegas</GEO> for fun.
```

Der Automat *married.grf* greift als erster von den hier vorgestellten Transduktoren auf den Graphen *rel_time_span.grf* zu. So würde sich dieser Subgraph auf das Identifizieren von relativen Zeitspannen konzentrieren. Das heißt nichts anderes, als dass Zeiträume im Korpus gesucht werden, die auf einen Zeitpunkt im Text Bezug nehmen, der eventuell das Erscheinungsdatum des Artikels ist, das aktuelle Jahr oder ein Datum, dass einige Passagen zuvor genannt wurde. Auf diese Weise würden auch Ausdrücke wie *5 years ago, 2 month later, days ahead, December ago, etc.* von der Grammatik erfasst werden.

10.5 Die Scheidung: „to be divorced“

Eine Heirat ist leider auch die Voraussetzung für eine Scheidung. Auch syntaktisch gesehen ähneln diese beiden Antonyme sehr, was die Grammatik in Abbildung 10.8 auf Seite 109 sehr deutlich aufzeigt.

So beschäftigt sie sich unter anderem mit der Feststellung nach wie vielen Ehejahren eine Beziehung geschieden wurde, wann dies geschehen ist, und eventuell noch wo die Scheidung stattgefunden hat.

Eine Konkordanz zum Graphen *divorced.grf* könnte beispielsweise wie folgt aussehen:

```
three months after Cristina had broken up with <Person><PersonName>
                        Phillip</PersonName></Person>
Today she is divorced and hoping to marry Yitzhak Rabin's assassin
                  bass player Dave Pegg is getting divorced
              John Hughes, 51, is divorced with three children.
              The Ohio congressman, who has been divorced twice
11 days after having been divorced, she married Captain Thomas George Symonds
                                    Babb
```

Um jedoch zu vermeiden, dass sich eine Firma von einem ihrer Mitarbeiter „scheiden“ lässt, wie es z.B. in folgendem Satz der Fall ist,

```
ICView -- Penna has parted company with its <Person><JD>previous chief
                executive</JD> and <JD>chairman</JD>
```

kommt der endliche Automat *merger_divorced.grf* zum Einsatz, welcher sicherstellt, dass nur Verbalphrasen mit menschlichen Nominalphrasen im Subjekt gefunden werden.

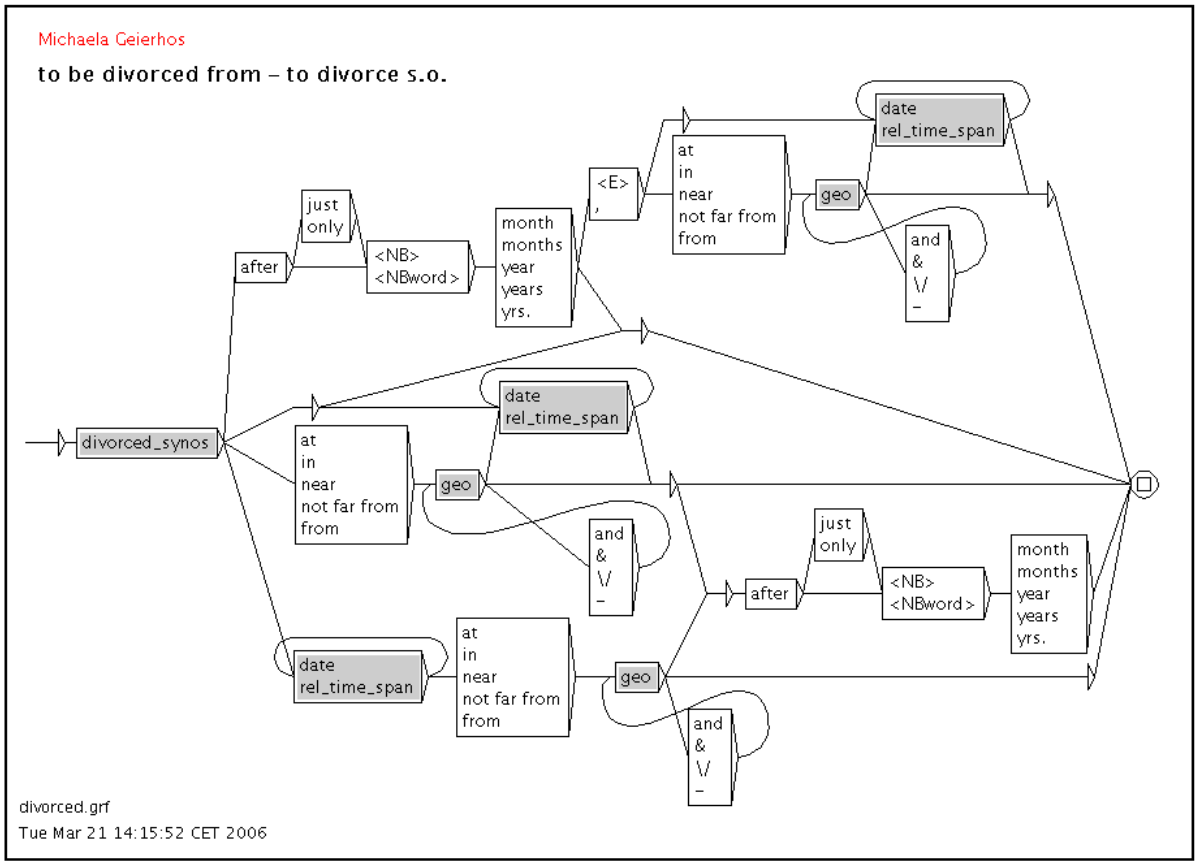


Abbildung 10.8: Graph zur Erkennung von Verbalphrasen mit dem Verb *to be divorced* und seiner direkten Synonyme – *divorced.grf*

Insgesamt werden 14 Verbkonstruktionen mittels dieser Grammatik im Korpus gefunden, welche im Subgraphen *divorced_synos.grf* (siehe Abbildung 10.9 auf Seite 110) ausführlich beschrieben werden. Dabei wird sowohl auf die Aktiv- als auch auf die Passivverwendung dieser Prädikate eingegangen.

- *to divorce so.*
- *to be divored from so.*
- *to file for a divorce from so.*
- *to sue for a divorce from so.*
- *to get a divorce from so.*
- *to part from so.*
- *to part company with so.*
- *to separate from so.*
- *to split from so.*
- *to split up with so.*
- *to break up with so.*
- *to end one's marriage to so.*
- *to dissolve one's marriage to so.*
- *to annul one's marriage to so.*

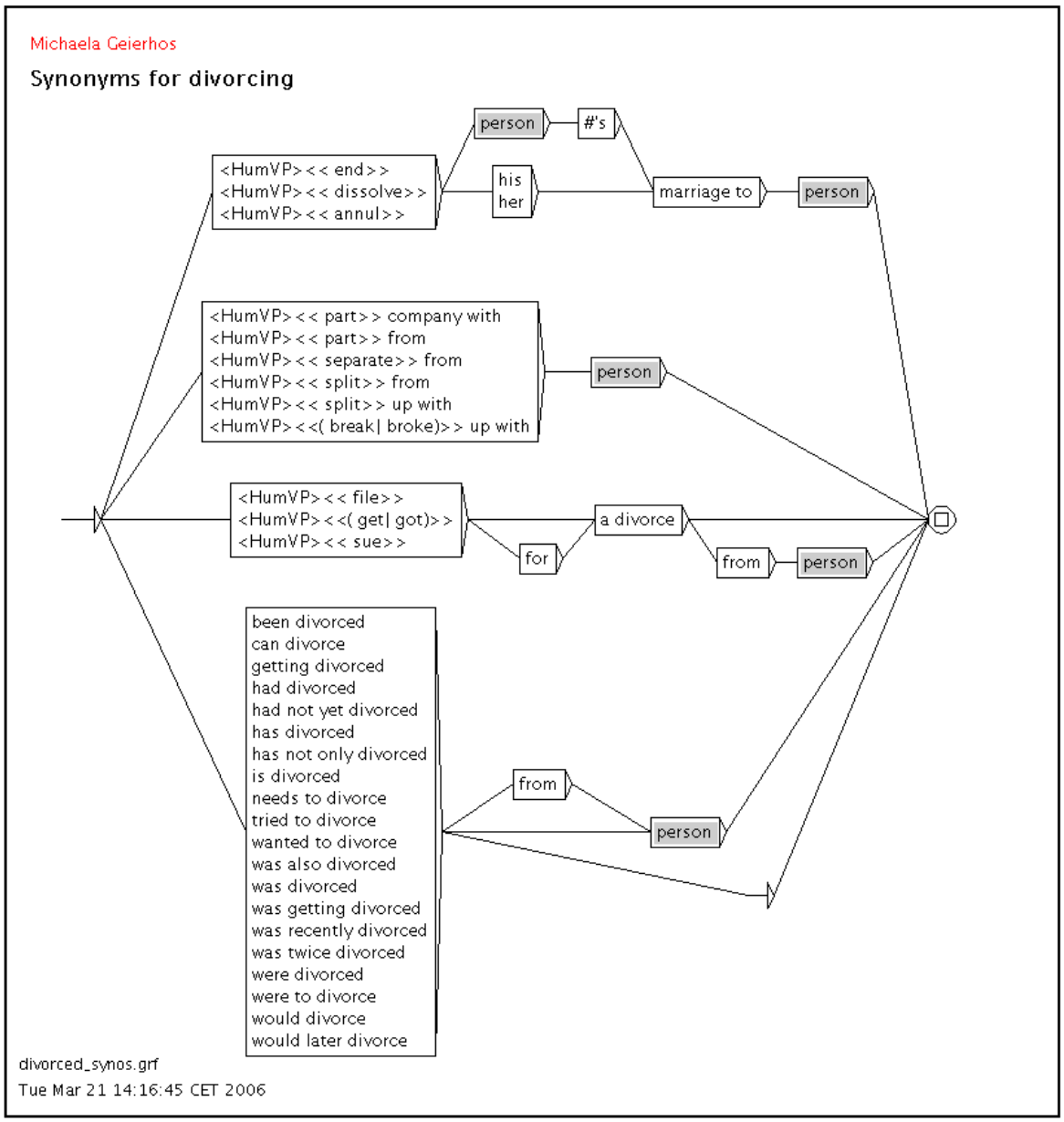


Abbildung 10.9: Graph zur Erkennung von Verben mit der Bedeutung von *„to be divorced“* – *divorced_synos.grf*

Überdies ist der Begriff der Synonymie hier relativ weit gefasst. Zwar entsprechen die meisten dieser Ausdrücke in ihrer Bedeutung einer Scheidung, doch können manche von ihnen auch nur eine Trennung oder das Ende einer Beziehung ohne Trauschein darstellen. Obwohl natürlich der Begriff der „Scheidung“ auch den Aspekt des „getrennt Lebens“ beinhaltet, sagt er doch noch etwas über die rechtliche Grundlage einer Partnerschaft aus. Doch dieser Faktor muss hier nicht berücksichtigt werden, und die syntaktische und semantische Ähnlichkeit der Verben reicht aus, sie in einem Graphen zusammenzufassen.

10.6 Der Tod: „to die“

Am Ende dieser Reihe von ausgewählten persönlichen Relationen stehen die Verben, welche den Tod eines Menschen in Worte fassen.

- *to breath one's last*
- *to decease*
- *to depart one's life*
- *to die (off)*
- *to expire*
- *to lay down one's life*
- *to lose one's life*
- *to meet one's death*
- *to meet one's end*
- *to pass away*
- *to perish*

Ausgehend von der folgenden Konkordanz, welche hauptsächlich nur darüber Aufschluss gibt, wann eine Person verstorben ist, oder wie viele Jahre ihr Tod schon zurückliegt, soll eine umfassende Grammatik entwickelt werden, welche weitere Informationen über den Tod des jeweiligen Menschen herausfindet.

Abbildung 10.10: Konkordanz zum Graphen *died.grf*

Doch gibt es noch andere Ergänzungen, welche im Umfeld des Ereignisses „Sterben“ auftreten. Diese Fülle an Fakten, welche mit dem Ableben einer Person einhergeht, versucht der Graph auf Seite 112 in den Griff zu bekommen.

Dabei sollte das Alter beim jeweiligen Todeszeitpunkt nicht außer Acht gelassen werden, welches meist in Phrasen des folgenden Typs ausgedrückt wird:

```
Prunella Clough died <Date>in 1999</Date> at the age of 80
        and died <Date>last month</Date> at age 83
Michael Dixon, who has died aged 71, was one of the Financial Times's longest-serving
                                columnists
Adolf Mahr died in <GEO>Bonn</GEO> in 1951, aged 64.
```

Ein weiterer – nicht unbedeutender – Faktor ist die Todesursache. Diese kann durchaus vielfältig sein, weil sie sich von diversen tödlichen Krankheiten und Drogenmissbrauch,

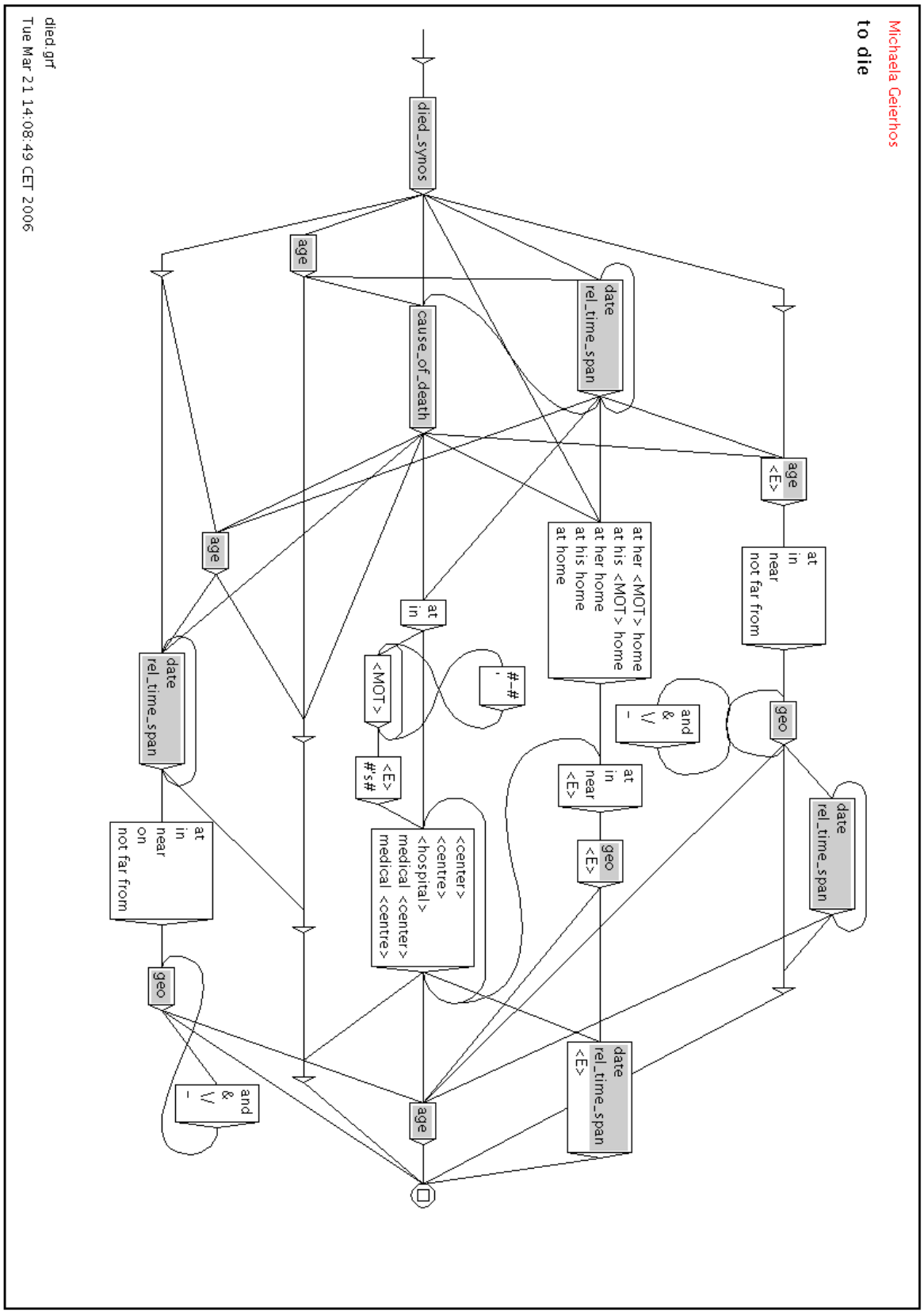


Abbildung 10.11: Graph zur Erkennung von Verbalphrasen mit dem Verb *to die* – *died.grf*

über die unterschiedlichsten Unfälle mit Todesfolge, bis hin zu simplem Herzversagen erstrecken kann.

Aus diesem Grund wurde ein eigener Graph erstellt, welcher nur die Aufgabe hat, Todesursachen im Korpus zu erkennen (siehe Abbildung 10.12).

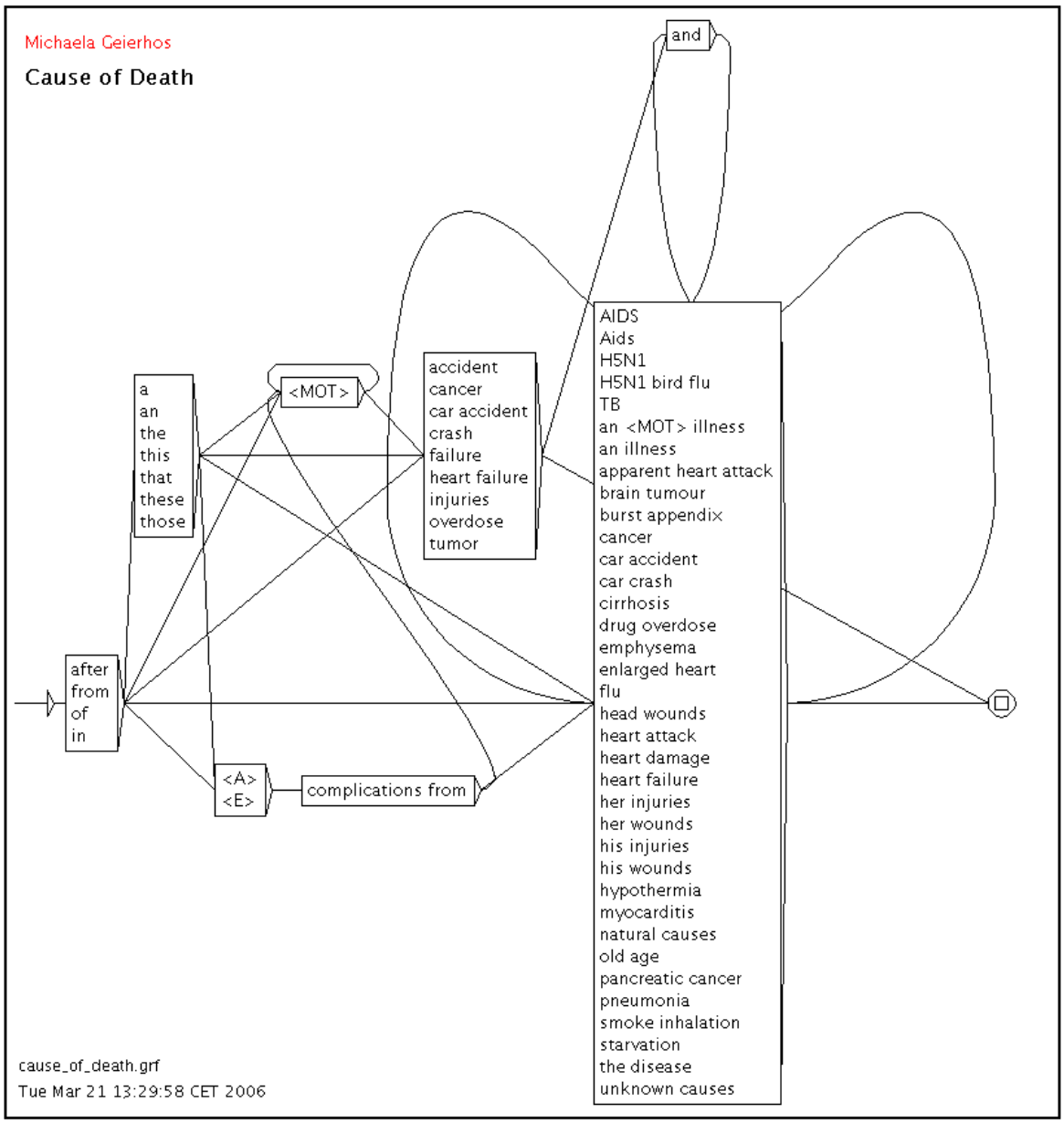


Abbildung 10.12: Graph zur Erkennung von möglichen Todesursachen – *cause_of_death.grf*

So würde der endliche Automat *cause_of_death.grf* beispielsweise folgende Todesursachen im Text aufspüren:

Thomas Hickey died in a cycling accident
Milt Bernhart, who has died of heart failure aged 77
Kristen Pfaff, had died of a heroin overdose

Zudem müssen in der Grammatik *died.grf* (siehe Seite 112) die verschiedenen Möglichkeiten erwähnt werden, welche ausdrücken, wo ein Mensch verstorben ist. Dabei sollte sie sich nicht nur auf Ortsangaben wie Städte oder Länder beschränken, sondern auch in Betracht ziehen, dass eine Person zu Hause oder im Krankenhaus sterben kann und dann keine genauere Ortsbestimmung darauf folgt.

```
Shirley Strickland de la Hunty has died at her <GEO>Perth</GEO> home aged 78
  Fischer died <Date>on Sunday</Date> in a hospital in <GEO>Lugano</GEO>
```

Der Subgraph *died_synos.grf* in Abbildung 10.13 ist das Kernstück des Hauptgraphen *died.grf*, denn er übernimmt die Beschreibung der einzelnen Verbkonstruktionen. Alle Prädikate, die in ihrer Bedeutung dem Verb *„to die“* entsprechen, kommen im Vokabular dieses Automaten vor.
So erfasst dieser Transduktor auch folgende Verbalphrasen:

```
Colonel Ryszard Kuklinski, 74, passed away in a <GEO>Washington</GEO> hospital
 Warren Zimmermann, who died of pancreatic cancer <DATE>last Tuesday</DATE>
```

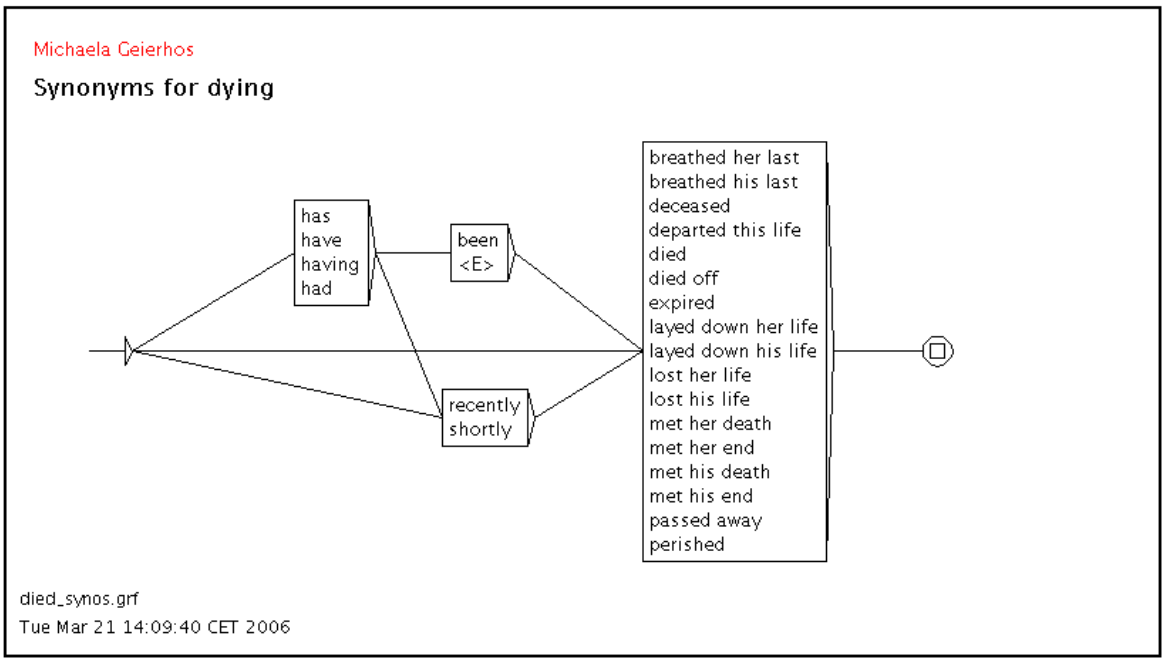


Abbildung 10.13: Graph zur Erkennung des Verbs *to die* und seiner direkten Synonyme – *died_synos.grf*

11 Grammatik beruflicher Relationen

Wie bereits zu Anfang in Abschnitt 1.2.2 angesprochen wurde, liegt der Schwerpunkt der hier vorgestellten linguistischen Untersuchungen verstärkt auf öffentlichen Relationen. Dabei stellen die beruflichen Relationen wohl die größte Untermenge der öffentlichen Beziehungen dar, denn die Möglichkeiten, ein Arbeitsverhältnis kombiniert mit einer Berufsbezeichnung wiederzugeben, sind vielfältig.

So versucht der endliche Automat aus Abbildung 11.1 ähnlich wie schon der Graph auf Seite 97 die einzelnen Grammatiken für ausgewählte Verbalphrasen auf recht kompakte Weise darzustellen. Auch hier werden Sätze mit derselben Struktur wie in der Grammatik für die persönlichen Relationen beschrieben.

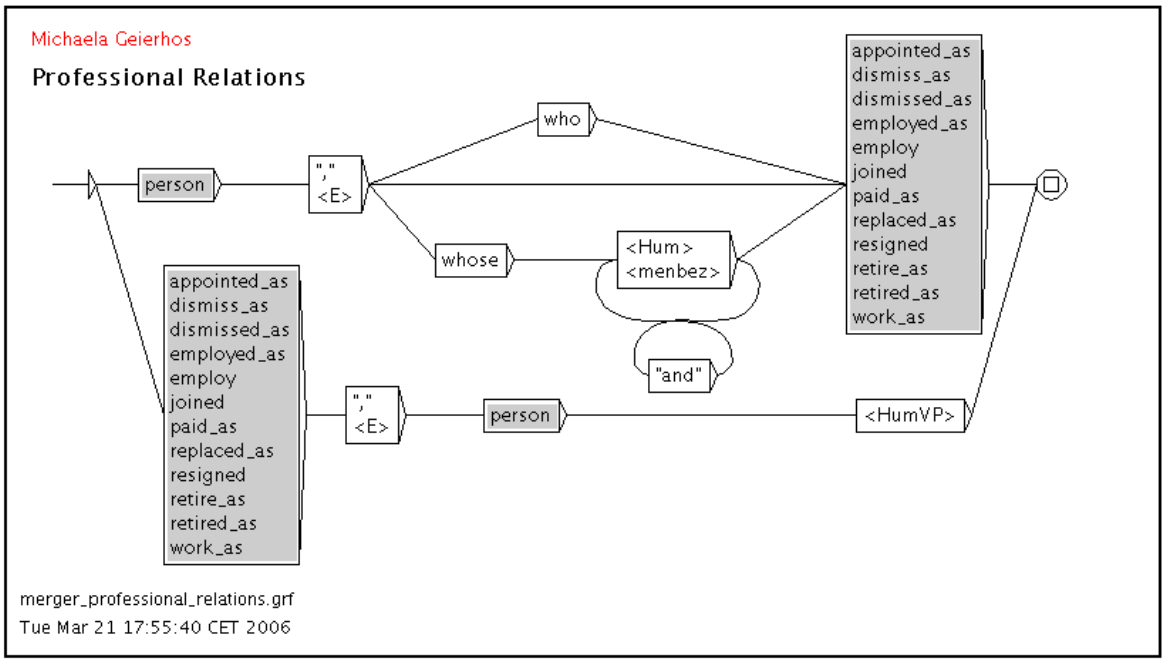


Abbildung 11.1: Graph zur Erkennung von ausgewählten beruflichen Relationen – *merger_professional_relations.grf*

Zudem wurde sehr auf die Modularität dieses Transitionsnetzes geachtet, so dass jeder einzelnen Relation ein eigener Graph zugewiesen wurde. Dabei decken die Subgraphen nicht nur die Verben ab, welche schon im Namen der Automaten vorkommen, sondern auch noch deren Synonyme. Auf diese Weise behandelt der Graph *merger_professional_relations.grf* 95 Verbkonstruktionen in verschiedenen Zeitformen und Variationen.

Mithilfe dieser Prädikate erfasst die Grammatik der beruflichen Relationen die wichtigsten Beziehungen, welche zwischen einer Person, ihrem Beruf und einer Firma be-

stehen können. Darunter fallen einerseits Verbalphrasen, die den Anfang eines Arbeitsverhältnisses oder den Beginn einer neuen beruflichen Karriere ausdrücken. Andererseits dürfen auch die Verben nicht fehlen, welche die Art einer Beschäftigung wiedergeben und diejenigen, die das Ende eines Beschäftigungsverhältnisses in Worte fassen. Des Weiteren besteht auch hier die Möglichkeit jeden der Subgraphen einzeln im Kontext des Graphen *merger_professional_relations.grf* (siehe Seite 115) aufzurufen, um sich ein Bild von der jeweiligen Relation im Satz zu verschaffen.

In den nächsten Abschnitten werden die verschiedenen Grammatiken präsentiert, welche die Verbalphrasen mit den bereits genannten Prädikaten beschreiben. Außerdem wurden im Zusammenhang mit den Verben entsprechende Automaten entwickelt, um auf die Struktur einer Berufsbezeichnung eingehen zu können.

11.1 Der Beginn eines Beschäftigungsverhältnisses

11.1.1 Die Ernennung: „to be appointed as“

Wenn jemand einmal eine gewisse Position in einem Unternehmen erlangt hat, so wird derjenige beispielsweise zum Vorsitzenden, Chef oder Direktor ernannt. Eine andere Möglichkeit ist die Ernennung zum Richter, Professor oder Kardinal, wobei diese Berufsgruppen in Wirtschaftsnachrichten nicht so zahlreich wie z.B. die Manager vertreten sind. Auch in der Politik werden gewählte Volksvertreter unter anderem als Gouverneur, Minister, Präsident usw. in ihr jeweiliges Amt eingeführt.

So bietet es sich an, Verben auszuwählen, welche den Aspekt der Berufung beinhalten, sich jedoch nicht auf eine bestimmte Domäne beschränken.

- *to be adopted*
- *to be appointed*
- *to be chosen*
- *to be commisssioned*
- *to be coopted*
- *to be designated*
- *to be elected*
- *to be engaged*
- *to be installed*
- *to be named*
- *to be nominated*
- *to be selected*
- *to be voted in*

Diese Passivkonstruktionen werden vom Graphen *appointed_synos.grf* behandelt, welcher als Subgraph in der Grammatik zu *„to be appointed as“* zum Einsatz kommt. Dabei ist die eigentliche Aufgabe der Grammatik aus Abbildung 11.2 die Erkennung von Verbalphrasen, in denen eine Person einen bestimmten Beruf ausübt oder eine gewisse Position erlangt. Im Umfeld von so genannten „Ernennungsrelationen“ sind hilfreiche Informationen zu finden, welche die Umstände dieser Amtseinsetzung näher spezifizieren. So wird meist im Zuge einer Amtseinführung angegeben, wessen Nachfolge nun angetreten wird. Um auf diese Textpassagen nicht verzichten zu müssen, wurde der Subgraph

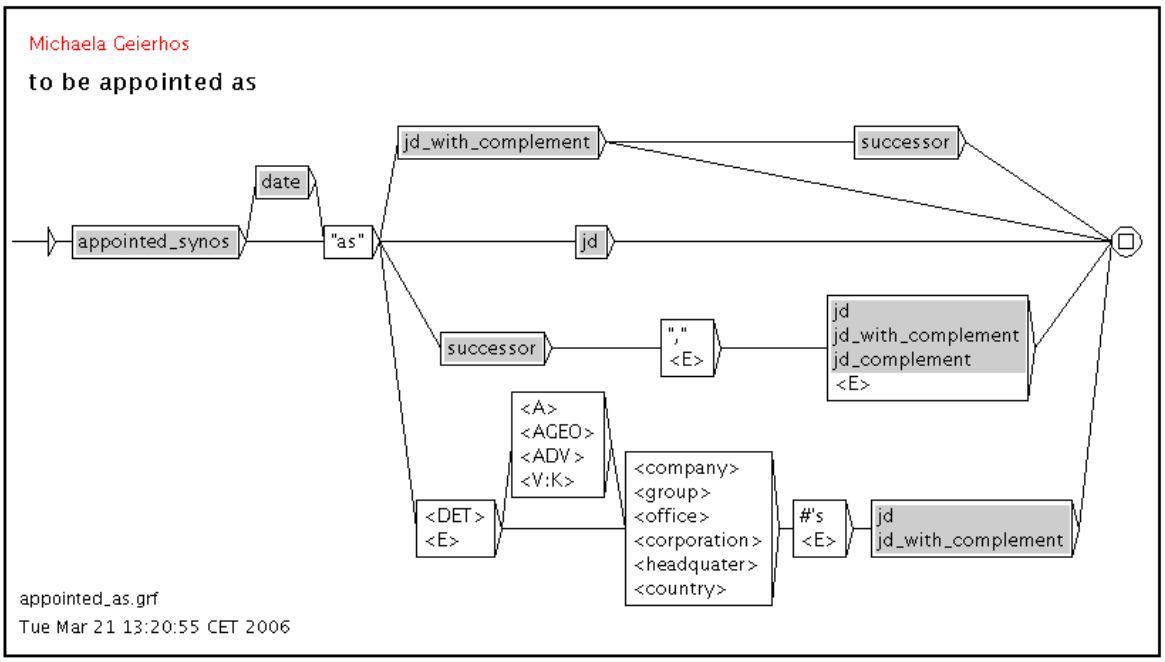


Abbildung 11.2: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to be appointed as“* und seiner direkten Synonyme – *appointed_as.grf*

successor.grf (siehe Abbildung 11.3 auf Seite 118) entwickelt, welcher Personennamen im Kontext einer „Nachfolgerelation“ aufspürt.

```
is expected to be appointed as successor to <JD>finance director</JD>
<PersonName>Manfred Gentz</PersonName>
Hart had been appointed as the <JD>Forest manager</JD> <Date>in July 2001</Date> as
successor to <PersonName>David Platt</PersonName>
Schily has appointed as <PersonName> Kersten</PersonName>'s successor
director of the Madrid-based offices, has been appointed as <PersonName>Mr
Herrero</PersonName>'s successor
is rumoured to be appointed as successor to <PersonName>Rolf Eckrodt</PersonName>,
<JD>chairman</JD> of <ORG>Japanese carmaker Mitsubishi</ORG>
```

Einen Vorgeschmack auf die volle Funktionalität des Graphen *appointed_as.grf* gibt die folgende Konkordanz aus dem FT-Korpus, wobei die Verberkennung vom Automaten *appointed_synos.grf* vorgenommen wurde:

```
Arif Khan had been appointed as <PersonName> Hizb-ul Mojahedin</PersonName>'s
<JD>divisional commander</JD> for <GEO>south Kashmir</GEO> two years ago
after being appointed as <GEO>Netanya</GEO>'s <JD>coach</JD> <Date>on Sunday</Date>
following <PersonName>Eli Cohen</PersonName>'s resignation
Tun Mohamed Dzaiddin Abdullah has been appointed as <JD>chairman</JD> of
<ORG>Deutsche Bank Malaysia</ORG>
Adrian Spencer Keane (41) has been appointed as <JD>Finance Director</JD> and
<JD>Company Secretary</JD> with effect from <Date>1 May 2004</Date>
Charles Craven has been appointed as <JD>director</JD> of <Sector>strategic
consulting</Sector>
Paolo Ceretti has been appointed as <JD>general manager</JD> of <ORG>Italian
publishing group De Agostini.</ORG>
```

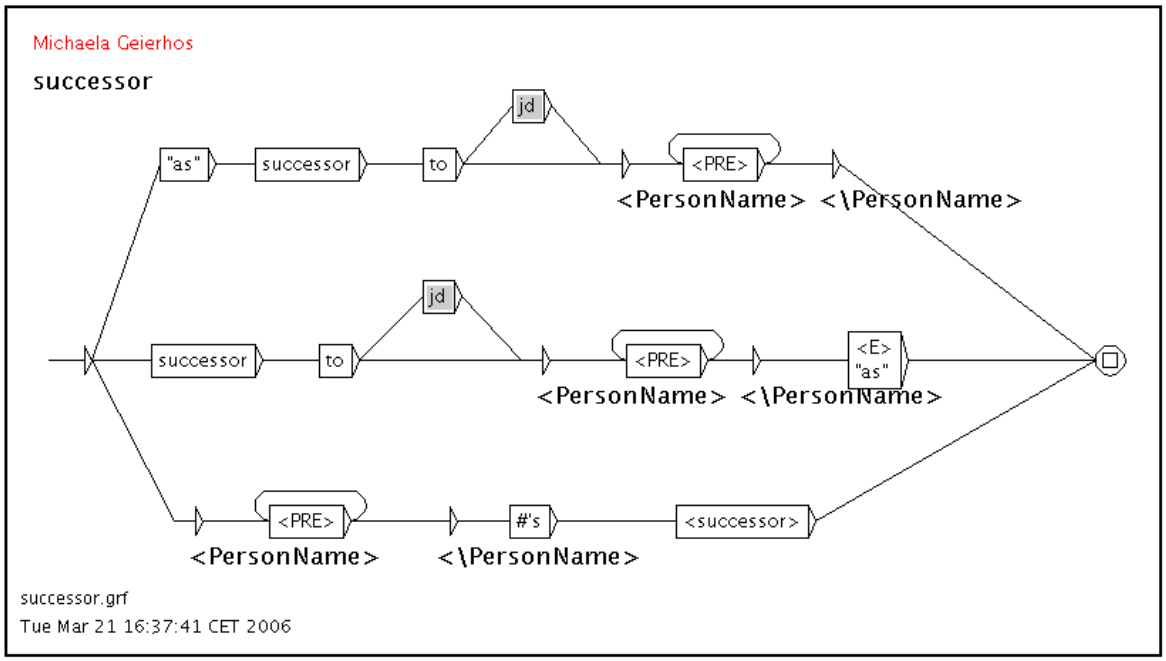


Abbildung 11.3: Graph zur Erkennung von Nachfolgern – *successor.grf*

Hierbei wurden Berufsbezeichnungen an den verschiedensten Positionen im Text erkannt. Wie der Automat diese identifizieren konnte, wird im nächsten Abschnitt ersichtlich.

Grammatik der Berufsbezeichnungen

Wie die Beispielkonkordanz zur Nachfolgerelation auf Seite 117 gezeigt hat, sind Berufsbezeichner zwar auch als Attribute von Personennamen im Korpus zu finden, doch sollten sie an erster Stelle die Position angeben, für welche die jeweilige Person vorgesehen ist.

Um die syntaktische Variabilität von Berufsbezeichnern in den Griff zu bekommen, wurde eine Reihe von Automaten entworfen, welche sich diesem Problem annehmen. In Abbildung 11.4 auf Seite 119 ist der Hauptgraph *jd.grf* zu sehen, welcher weitere Subgraphen koordiniert. Dieser Transduktor stützt sich nicht nur auf das in den Lexika kodierte Wissen über Berufe, sondern versucht auch den linken Kontext von Berufsbezeichnungen genauer zu beschreiben. Dabei geht er auf mögliche Adjektive bzw. Adverbien, sowie auf Nominalphrasen ein, welche die Beschäftigung einleiten und näher spezifizieren.

Phrasen, in denen diese Adjektive wichtige Eigenschaften über die Arbeit der betreffenden Person aussagen, könnten wie folgt aussehen:

```
Deputy Prime Minister Viktor Khristenko has been appointed as <JD>acting prime
                                    minister</JD>
   is expected to be appointed as <JD>general production co-ordinator</JD>
the finance director of Pearson PLC, has been appointed as a <JD>non-executive
                                    director</JD>
```

Um diese attributiven Ergänzungen zu den jeweiligen Berufsbezeichnungen zu lokalisie-

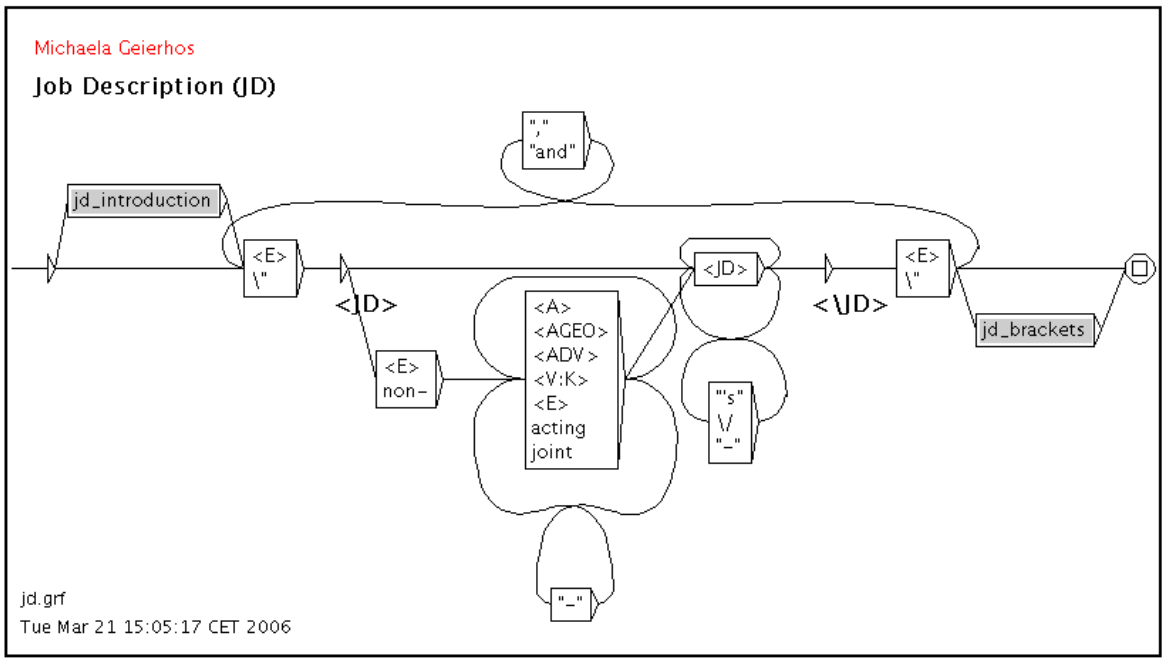


Abbildung 11.4: Graph zur Erkennung von Berufsbezeichnungen – *jd.grf*

ren, reichte es über die Symbole `<A>`, `<AGEO>`, `<ADV>` und `<V:K>` in den entsprechenden Lexika nachzuschlagen und auf diverse Schlüsselbegriffe wie *„acting“* oder *„joint“* zurückzugreifen.

Dagegen ist es für die einleitenden Nominalphrasen notwendig, eine Grammatik anzugeben. Sie hat dann die Aufgabe Phrasen wie diese

```
after being appointed as <GEO>Netanya</GEO>'s <JD>coach</JD>
Arif Khan had been appointed as <PersonName>Hizb-ul Mojahedin</PersonName>'s
                        <JD>divisional commander</JD>
Bulloch has been installed as <GEO>Paterson</GEO>'s <JD>vice-captain</JD>
Craig Moore has been named as <GEO>Australia</GEO>'s <JD>captain</JD>
a former police chief constable, was appointed as <PersonName>Tony
Blair</PersonName>'s <JD>chief drugs fighter</JD>
```

im Korpus zu erkennen.

Wie bereits in Abbildung 11.2 auf Seite 117 ersichtlich war, gibt es noch zwei weitere Grammatiken, welche den rechten Kontext des Verbes *„to be appointed as“*, sowie seiner Synonyme beschreiben. Dabei handelt es sich um den Automaten *jd_complement.grf*, welcher den allgemeinen rechten Kontext einer Berufsbezeichnung spezifiziert und *jd_with_complement.grf*, der die Berufsbezeichnung gemeinsam mit ihrem attributiven rechten Kontext wiedergibt. Die beiden Graphen werden auf den Seiten 120 und 121 dargestellt. Dabei ist der Inhalt der Subgraphen fast selbsterklärend, da sie entweder Gründe (*„dismissal“*, *„resignation“*) angeben, warum der Posten neu besetzt wurde, oder ab wann die betreffende Person ihr Amt inne hat, oder ob die Ernennung eventuell auf der Jahresversammlung einer Firma bekannt gegeben wurde.

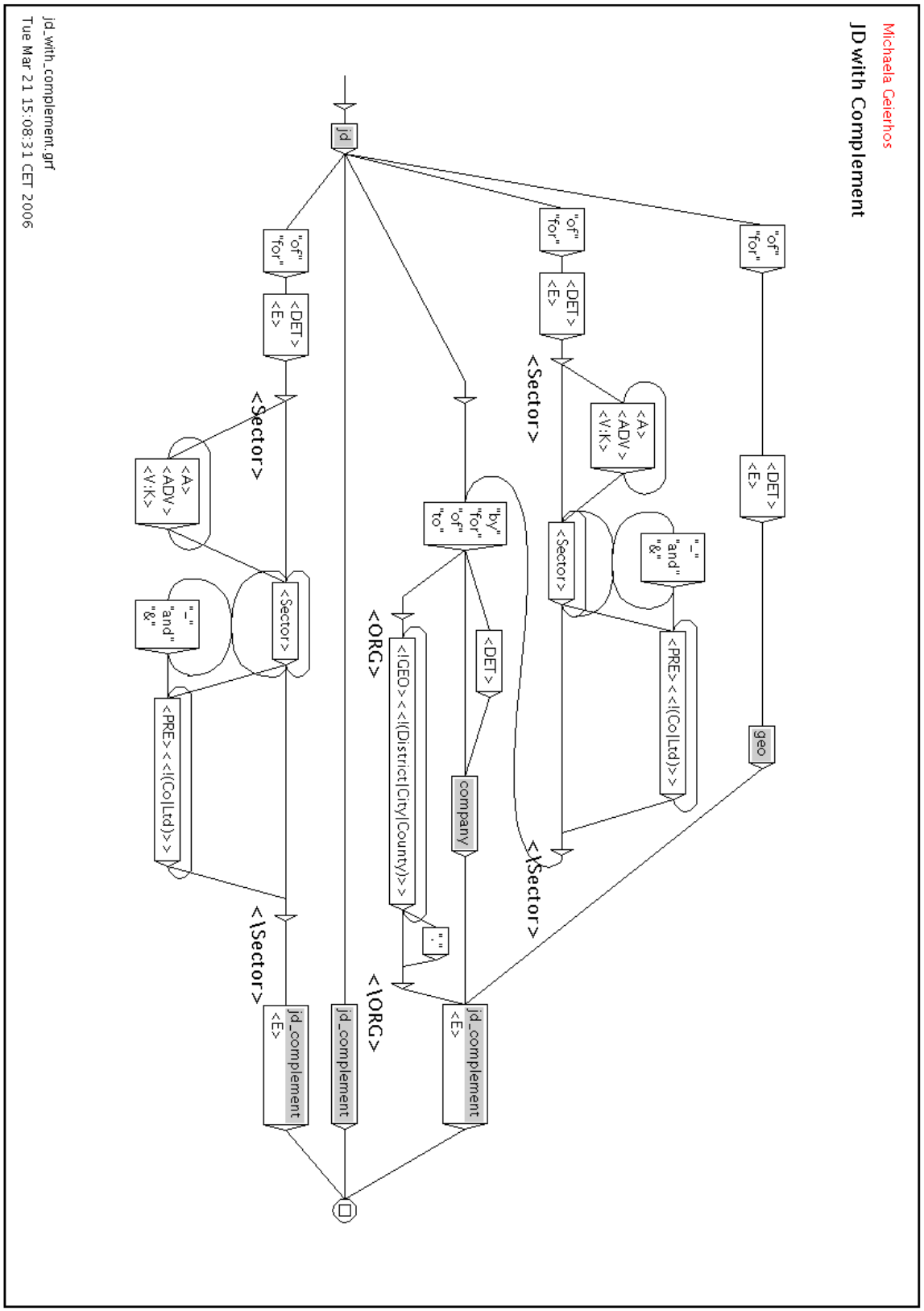


Abbildung 11.5: Graph zur Erkennung von Berufsbezeichnern mitsamt ihrer rechten Kontexte – *jd_with_complement.grf*

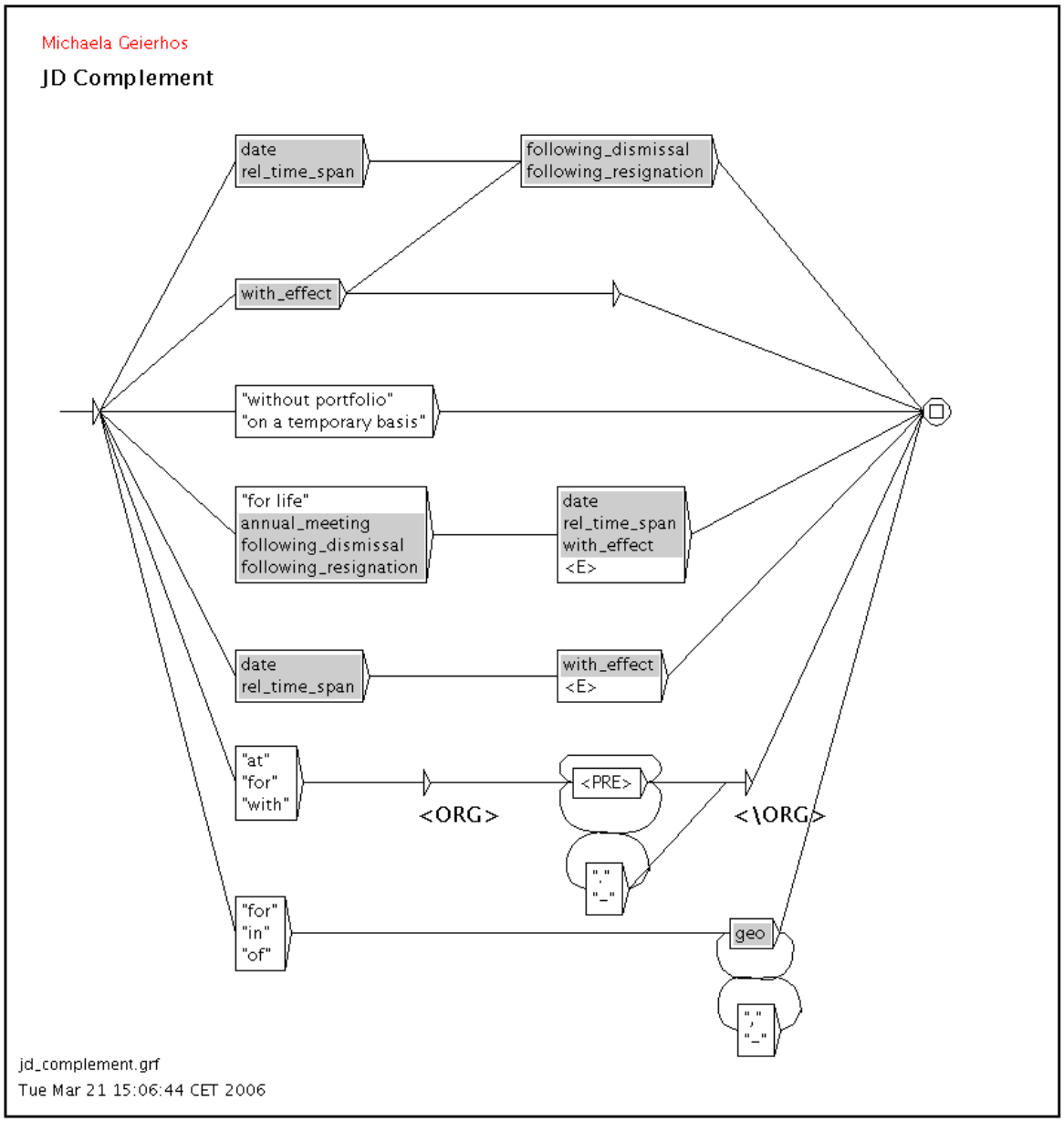


Abbildung 11.6: Graph zur Erkennung der Komplemente von Berufsbezeichnungen (rechter Kontext eines Berufsbezeichners) – *jd_complement.grf*

Vervollständigung des Sektorenlexikons

Bereits im Graphen *jd_with_complement.grf* auf Seite 120 fiel der Begriff des Sektors, der dort als Lexikonreferenz in Form von `<Sector>` eingesetzt, und später in der Ausgabe des Transduktors entsprechend annotiert wurde. In diesem Kontext wurden keine Variablen anstelle des Symbols `<Sector>` zugelassen, weil die vorangehende Präposition oft nicht ausreicht um eine Branche eindeutig zu identifizieren. Da manche Firmen Begriffe wie *„Marketing“*, *„Design“* oder *„Food“* enthalten, könnten Firmennamen irrtümlicherweise zu Sektorennamen und einem unbedeutenden Rest aufgespalten werden.

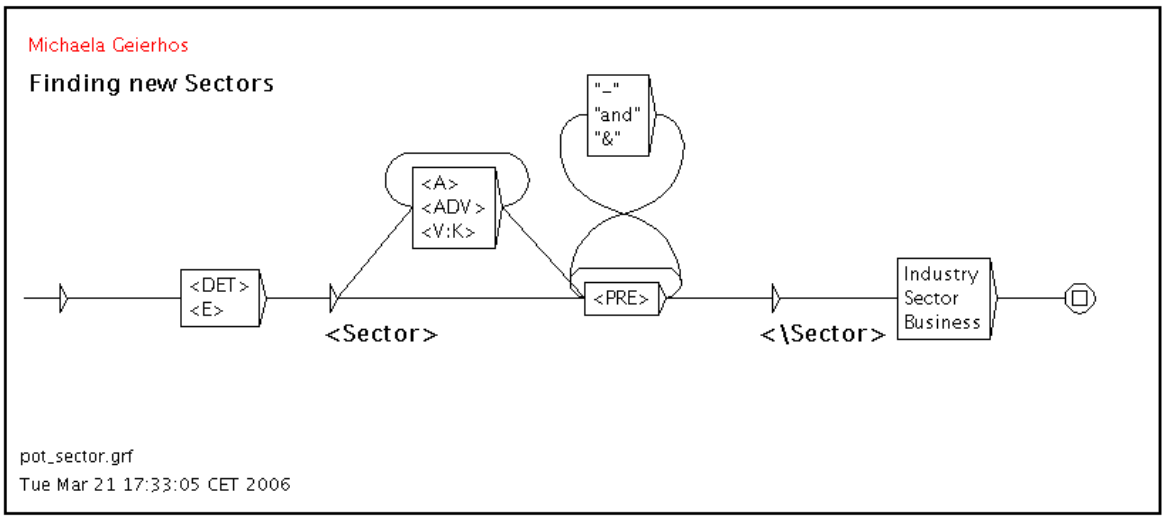


Abbildung 11.7: Graph zur Erkennung potentieller Sektoren- und Branchenbezeichnungen – *pot_sector.grf*

Aus diesem Grund wurde der endliche Automat aus Abbildung 11.7 erstellt, welcher auf Schlüsselbegriffe im rechten Umfeld von potentiellen Branchenbegriffen baut, um qualitativ gute Ergebnisse bei der Suche nach neuen Wörterbucheinträgen zu erzielen.

Dieser Graph findet nun folgende Kandidaten für neue Branchenbegriffe, die noch nicht in Lexikon *Sector-.dic* (siehe Abschnitt 5.2.7 auf Seite 57) enthalten sind.

```
<Sector>Automated Applicant Screening and Scheduling Processes</Sector> Business
<Sector>Access & Networking</Sector> Business
<Sector>Communications</Sector> Business
<Sector>Enterprise Software & Storage</Sector> Business
<Sector>Industrial Applications</Sector> Business
the <Sector> Leather</Sector> Industry
the <Sector> Web Analytics</Sector> Business
```

Natürlich hätte man auch andere Indikatoren, wie z.B. *„Director of“*, verwenden können, um neue Sektorennamen zu erhalten. Doch wenn nicht mehr Kontext spezifiziert wird, sind beide Ansätze für Fehler äußerst anfällig und bedürfen einer manuellen Korrektur, bevor Begriffe automatisch aus dieser Liste extrahiert werden.

11.1.2 Die Einstellung: „to employ so.“

Nachdem die Verbkonstruktion *„to be appointed“* hier sehr ausführlich behandelt wurde, müssen die nachfolgenden Prädikate nicht mehr so intensiv behandelt werden. Das liegt unter anderem auch daran, dass sie alle auf die Berufsbezeichnergraphen (siehe Abschnitt 11.1.1) zugreifen.

So fällt beispielsweise die Tatsache, dass eine Firma oder eine Person jemanden einstellt, der für sie arbeitet, auch in die Kategorie der beruflichen Relationen. Hierbei handelt es sich sogar um eine Prädikatsbeziehung, die zwei Menschen betreffen kann, und ist somit für diese Arbeit von besonderem Interesse.

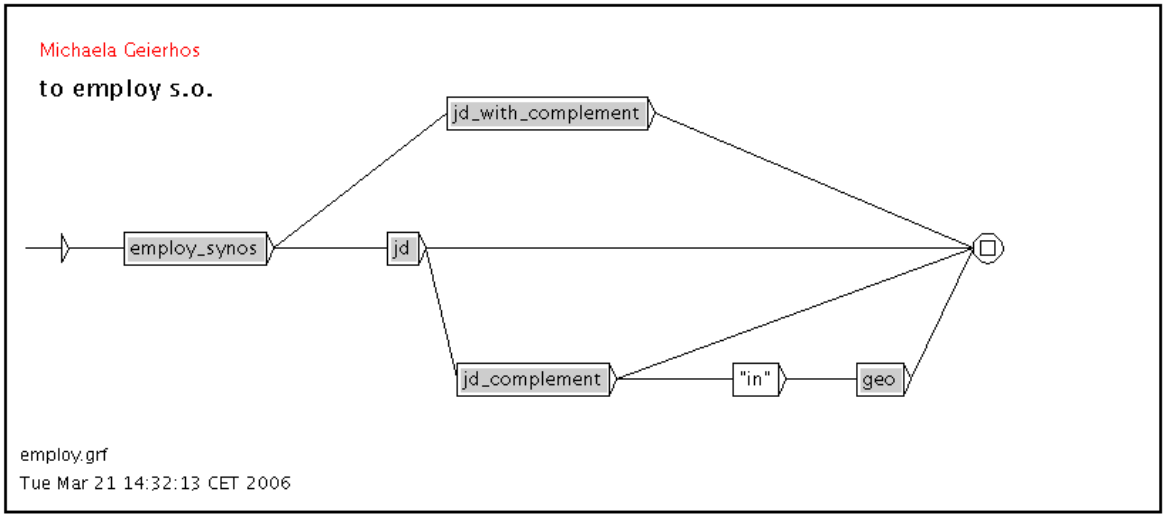


Abbildung 11.8: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to employ s.o.“* und seiner direkten Synonyme – *employ.grf*

Wenn man den Automaten in Abbildung 11.8 betrachtet, ist es leicht mit dem Vorwissen aus dem letzten Abschnitt seine Funktionsweise zu verstehen. Welche Verben er jedoch behandelt, wird nicht auf den ersten Blick ersichtlich, denn dafür ist der Subgraph *employ_synos.grf* zuständig, welcher folgende Verbkonstruktionen umfasst:

- *to enrol so.*
- *to employ so.*
- *to engage so.*
- *to enlist so.*
- *to hire so.*
- *to put so. on the payroll*
- *to recruit so.*
- *to sign (up) so.*
- *to take so. into employment*
- *to take on*
- *to retain so.*
- *to secure the services of so.*

Eine entsprechende Konkordanz zur Grammatik *employ.grf* (siehe Abbildung 11.8) könnte beispielsweise so aussehen:

```
it had hired <Person> <PersonName>Stewart Lawrie</PersonName> </Person> as
<JD>director</JD> of UK and international accounts from Proctor & Gamble
announced that it has hired <Person> <PersonName>Frank Fiorillo</PersonName>
</Person> as <JD>vice president</JD> of worldwide customer support
is pleased to announce that it has hired <Person> <PersonName>Jeri
Silverman</PersonName> </Person> as a <JD>consultant</JD>
Evergreen Investments has hired <Person> <PersonName>Louis Membrino</PersonName>
</Person> as a <JD>Director</JD> in Evergreen Consultant Relations
Mr Stewart has recruited <Person> <PersonName>Lynne Peacock</PersonName> </Person> as
<JD>business development director</JD> at NAB Europe
```

11.1.3 Der Firmeneintritt: „to join“

Im Englischen sind wohl die folgenden Verbkonstruktionen die gebräuchlichsten Möglichkeiten, um den Eintritt einer Person in eine Firma zu beschreiben:

- *to join*
- *to become a member of*

Aufgrund ihrer minimalen Anzahl werden sie direkt in der Grammatik *join.grf* behandelt. Diese lokalisiert unter anderem die jeweiligen Unternehmen, zu denen die betreffende Person gegangen ist. Hierfür wird der Automat *company.grf* (siehe Seite 80) eingesetzt, welcher Firmennamen auf sehr umfassende Weise beschreibt. Außerdem werden noch weitere Informationen, wie der Sitz der Firma mithilfe des Graphen *geo.grf* (siehe Seite 87), das Einstellungsdatum mit *date.grf* (siehe Seite 90), oder der Beruf der betreffenden Person (siehe *jd.grf* auf Seite 119), sowie Beschäftigungszeit durch den Automaten *rel_time_span.grf* ermittelt.

Der Transduktor *join.grf* würde die Verbalphrasen wie folgt annotieren:

```
Jim Sweeney will also be joining <ORG>AmeriQuest</ORG> as <JD>Vice President</JD>
Andrew Bird will be joining <ORG>Hot Group</ORG> as <JD>head</JD> of e-marketing
     is joining the <ORG>Selfridges board</ORG> as <JD>deputy chairman</JD>
         and join <ORG>Reuters</ORG> as <JD>non-executive chairman</JD>
       But Mr Stewart, who joined <ORG>BSkyB</ORG> <Date> in 1996</Date>
He joined <ORG>Caledonian Insurance Services Limited</ORG> <Date>in November</Date>
Trevor Williams FCIS (42) joined <ORG>Imperial Tobacco</ORG> <Date>in 1996</Date>
     Mr Rothschild, who joined <ORG>Tower Hotels</ORG> <Date>in 1992</Date>
```

Möchte man jedoch erfahren, wie der Name der Person lautet, welche in die mit `<ORG>` markierte Firma eingetreten ist, muss der Graph *merger_join.grf* (ohne Abbildung) aufgerufen werden, welcher dafür zuständig ist, ebenfalls das menschliche Subjekt im Satz zu kennzeichnen.

11.2 Die Ausübung des Berufes

Wurde eine Arbeitsstelle einmal angetreten, beginnt in der Regel die Zeit, in der ein Beruf ausgeübt wird. Die Tatsache, dass eine Person bei einem Unternehmen angestellt ist und somit dort arbeitet, kann auf die verschiedensten Arten ausgedrückt werden. Dabei muss es sich nicht nur um Aktiväußerungen handeln, in denen auf die Art der Tätigkeit des Arbeitnehmers eingegangen wird, es können auch Passivkonstruktionen sein, welche das aktuelle Beschäftigungsverhältnis eines Menschen beschreiben.

11.2.1 Das Beschäftigungsverhältnis: „to be employed“

Das Prädikat *„to be employed“* ist ein typisches Beispiel für so eine Passivphrase. Bereits in Abschnitt 11.1.2 wurde das Verb *„to employ“* in seiner Aktivform eingeführt. Dabei ging es um den Akt des Einstellens – eine Firma oder eine Person stellt einen Arbeitssuchenden bei sich ein. Bei dieser Variante handelte es sich um den Beginn eines

Arbeitsverhältnisses, wobei die passive Ausdrucksweise die Tatsache wiedergibt, dass jemand gerade bei einer Organisation beschäftigt ist. Das heißt nichts anderes, als dass derjenige schon die Phase des Firmeneintritts hinter sich gebracht hat und nun voll und ganz im Arbeitsleben steht.

Deshalb mussten lediglich die Verben des Aktivgraphen *„to employ“* für diese Grammatik in ihre jeweilige Passivkonstruktion überführt werden und konnten so im Automaten *employed_synos.grf* kodiert und von *employed_as.grf* verwendet werden.

- *to be enrolled*
- *to be employed*
- *to be engaged*
- *to be enlisted*
- *to be hired*
- *to be put on the payroll*
- *to be recruited*
- *to be signed (up)*
- *to be taken into employment*
- *to be taken on*
- *to be retained*
- *to have an employment*

Eine mögliche Konkordanz des Automaten *employed_as.grf* würde folgendermaßen aussehen:

was to be retained as a <JD>consultant</JD> to promote the scheme
his mother might have been employed as a <JD>domestic servant</JD>.
chief client officer Lynne Seid has been hired as <JD>president</JD> of FCB New York
Industry veteran Gabriel Pulido has been hired as <ORG> Captiva</ORG>'s <JD>business development manager</JD> for Central and South America
agent Michael Ovitz would never have been hired as <PersonName>Eisner</PersonName>'s <JD>next heir apparent</JD>
has been retained as the <JD>collateral manager</JD> of Independence
manager Humphrey Kelleher appears to have secured the services of <JD>dual player</JD> Conal Keaney for at least the remainder of the National Hurling League.
Mr Mills' former boss, is engaged as a <JD>consultant</JD> to Lonmin.
he was earlier employed as <JD>Chief Geologist</JD> of Giant Mascot Mine Ltd
Deziel was employed as a <JD>customer relationship marketing director</JD> at Philip Morris USA in New York City.
in 1992 was hired as <JD>chancellor</JD> of Seattle Community Colleges

11.2.2 Die Bezahlung: „to be paid as“

Wenn jemand in einem Unternehmen beschäftigt ist, wird derjenige in der Regel auch für seine Arbeit finanziell entschädigt. Also kann davon ausgegangen werden, dass die Bezahlung als Gegenleistung für eine bereits verrichtete bzw. noch fortführende Beschäftigung erfolgt. Diese Tatsache wird in der Regel im Passiv ausgedrückt, wobei der Name des Arbeitgebers meist in diesem Zusammenhang nicht fällt.

Um die Relation angemessen beschreiben zu können, wurden die folgenden beiden Prädikate dafür in Betracht gezogen.

- *to draw salary*
- *to be paid*

Im Graph der Abbildung 11.9 wurde nur die Präposition *„as“* im Anschluss an das jeweilige Verb zugelassen, auf welche eine Berufsbezeichnung folgen muss. Somit sollten nur Sätze in der Konkordanz enthalten sein, welche inhaltlich darauf Bezug nehmen, dass eine Person für eine bestimmte Tätigkeit (ihren Beruf) entlohnt wird.

```
Corman, who is paid as a <JD>consultant</JD>, holds 650,000 options
```

Dabei wird die Suche nach den entsprechenden Berufsbezeichnern und ihren Kontexten wieder von den Transduktoren *jd.grf* (siehe Seite 119), *jd_with_complement.grf* (siehe Seite 120) und *jd_complement.grf* (siehe Seite 121) übernommen. Außerdem werden bei der Floskel *„to draw salary as“* noch weitere Variationen zugelassen, indem noch zusätzlich die Begriffe

„wages, pay, earnings, fee, fees“

als Synonyme von *„salary“* im Vokabular des Graphen zugelassen werden.

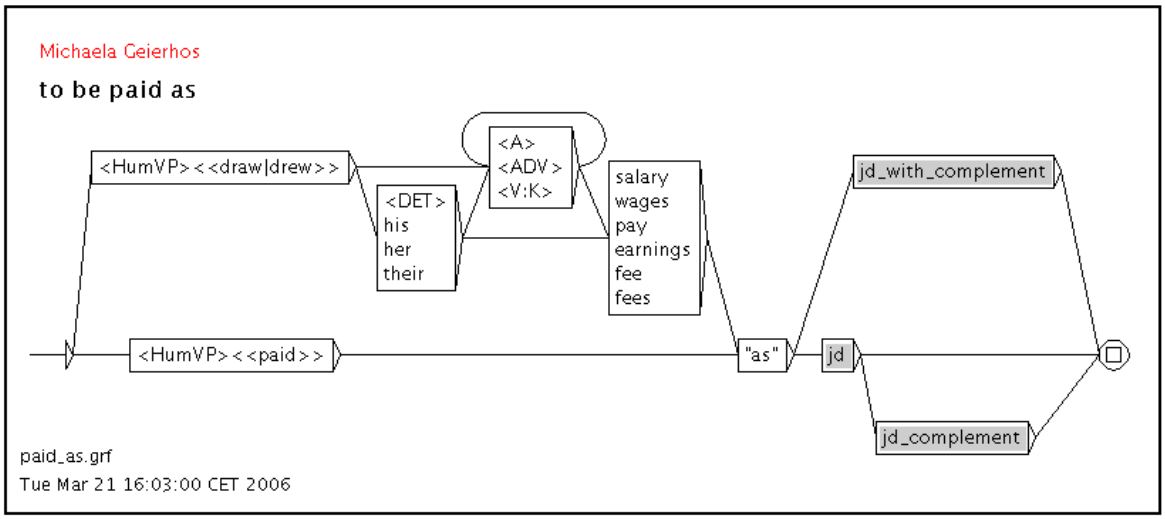


Abbildung 11.9: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to be paid as“* und seiner Paraphrasierung *„to draw salary as“* – *paid_as.grf*

11.2.3 Die Tätigkeit: „to work as“

Wenn es darum geht, das Betätigungsfeld einer Person näher zu spezifizieren, bietet sich an, einfach auf die Relation *„to work“* zurückzugreifen. Auch ist offensichtlich, dass sich diese Person in einem momentanen Beschäftigungsverhältnis befinden muss. Natürlich gibt es weitere Verben, welche den gleichen Sachverhalt ausdrücken:

- *to work*
- *to labour*
- *to job*
- *to labor*

- *to operate*
- *to serve*
- *to toil*

Zudem werden im Umfeld einer Arbeitsbeziehung wichtige Fragen geklärt, die sowohl den Arbeitgeber als auch den Arbeitnehmer betreffen:

1. Für wen arbeitet die betreffende Person? Wer ist ihr Arbeitgeber?
   a) Handelt es sich hierbei um eine Organisation?
   b) Oder wird sie von einer Privatperson beschäftigt?
2. Seit wann ist die betreffende Person für ihren Arbeitgeber tätig?
3. Wie lange arbeitet die betreffende Person schon bei diesem Arbeitgeber?
4. Von wann bis wann hat die betreffende Person dort gearbeitet?
5. Als was ist die betreffende Person dort beschäftigt? Welche Position hat sie inne?
6. In welcher Branche hat die betreffende Person gearbeitet?
7. Wie ist die betreffende Person beschäftigt? Vollzeit? Teilzeit?
8. Welches Alter hatte die betreffende Person, als sie dieser Tätigkeit nachging?

Die dazugehörige Konkordanz illustriert, welche Informationen der Transduktor im Text erkannt und annotiert hat.

```
Chief Operating Officer, has agreed to serve as <JD>interim Director</JD> of Network
                                                                    Marketing.
        President Jean-Bertrand Aristide can continue to serve effectively as
                                    <GEO>Haiti</GEO>'s <JD>leader</JD>.
   Gary L. Primes, who continues to serve as <JD>CIO</JD> in addition to his recent
                                                                  appointments as COO
and hired Gutman, who had been working as <JD>CEO</JD> of 300-employee company Liraz
     John Blue, who had served as <JD>Acting CEO</JD> from <Date>March 2003</Date>
David Brocklehurst, had served as <JD>finance director</JD> for several ad agencies
  who had served as <JD>justice secretary</JD> during the administration of deposed
                                                                  President Joseph
      bestselling author and historian who had served as <JD>librarian</JD> of
                                                                  <ORG>Congress</ORG>
        had served as <JD>police chief</JD> for <GEO>the Kushiro region</GEO>
  John Lewis Ashcroft had worked as a <JD>professional trapper</JD> near Guyra, at
                                                                  Kangaroo Camp.
    Wells Rich Greene, and has also served as <JD>Assistant Vice President</JD>,
                                                                  Marketing Communications
       He has also served as <JD>President</JD> of <ORG>Midway Airlines</ORG>
Col Mohammad Esa has been working as the <JD>acting head</JD> of the department.
Schneider has served as <JD>chief financial officer</JD> and <JD>principal</JD> of
                                    <ORG>Leonard Green & Partners</ORG>
  Sam Skinner has served as <JD>Co-Chairman</JD> of <ORG>Hopkins and Sutter</ORG>
```

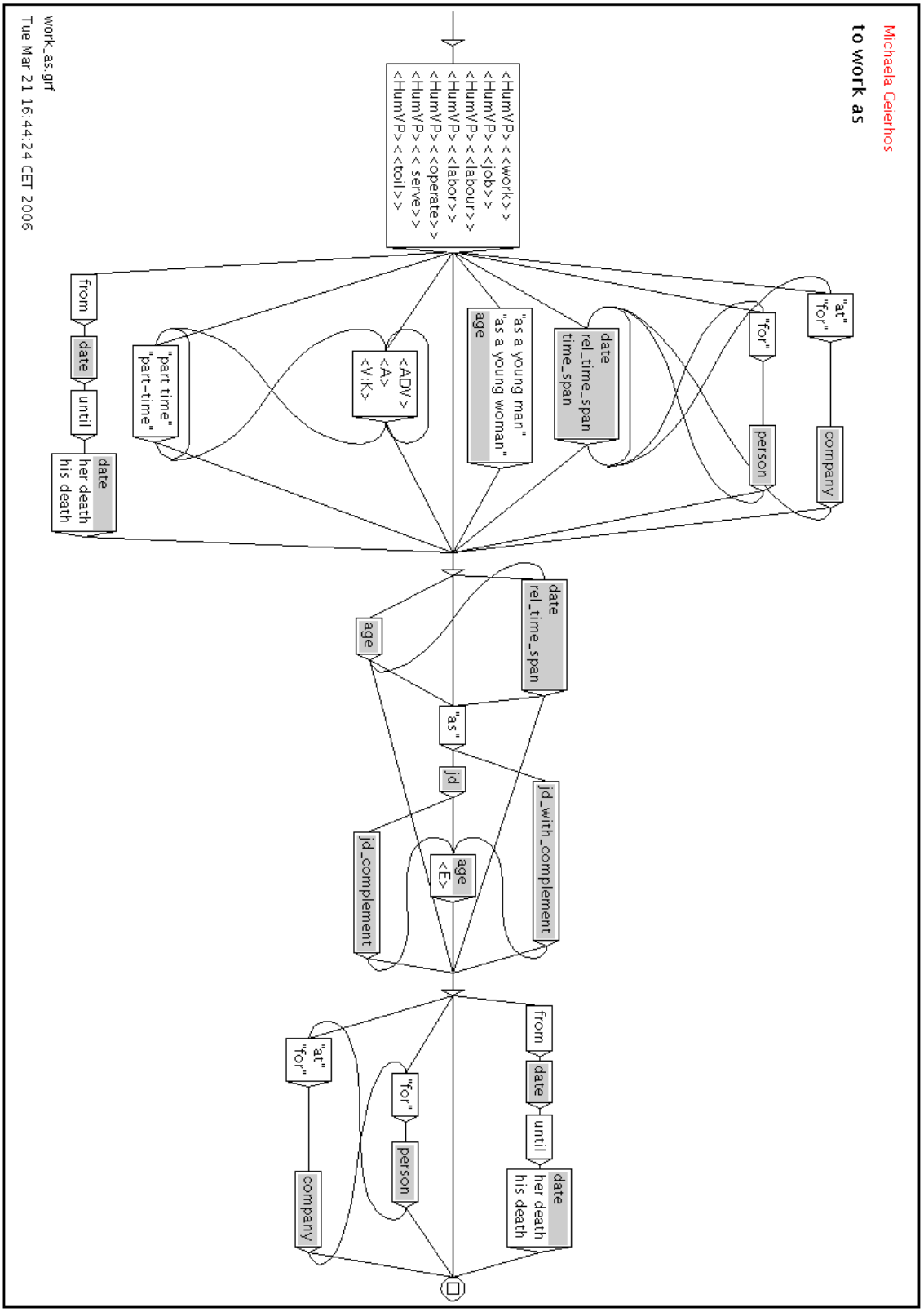


Abbildung 11.10: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to work as“* und seiner direkten Synonyme – *work_as.grf*

11.3 Das Ende eines Arbeitsverhältnisses

Jedes Arbeitsverhältnis wird irgendwann aufgelöst. Dabei kann es sich um ein abruptes Ende handeln, welches beispielsweise durch eine Kündigung ausgelöst wird, oder man ist sich des Termins bewusst, weil das Rentenalter erreicht wurde, ein besseres Jobangebot vorliegt, oder man freiwillig aus dem Berufsleben ausscheiden will.

11.3.1 Die Entlassung: „to dismiss so.“ bzw. „to be dismissed“

Im Englischen bringen folgende Verbkonstruktionen zum Ausdruck, dass ein Arbeitsverhältnis gekündigt wurde:

- *to decapitate so.*
- *to decruit so.*
- *to disband so.*
- *to discard so.*
- *to discharge so.*
- *to dismiss so.*
- *to displace so.*
- *to eject so.*
- *to expel so.*
- *to fire so.*
- *to give so. one's notice*
- *to lay (off)*
- *to make so. redundant*
- *to oust so.*
- *to pay so. off*
- *to release so.*
- *to relieve so.*
- *to remove so.*
- *to throw so. out*

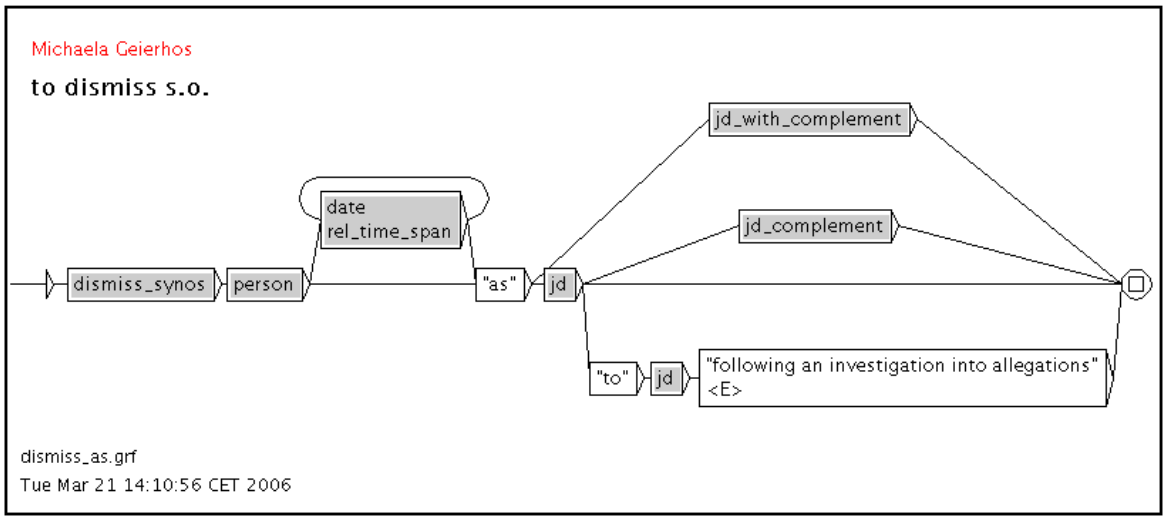


Abbildung 11.11: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to dismiss s.o.“* und seiner direkten Synonyme – *dismiss_as.grf*

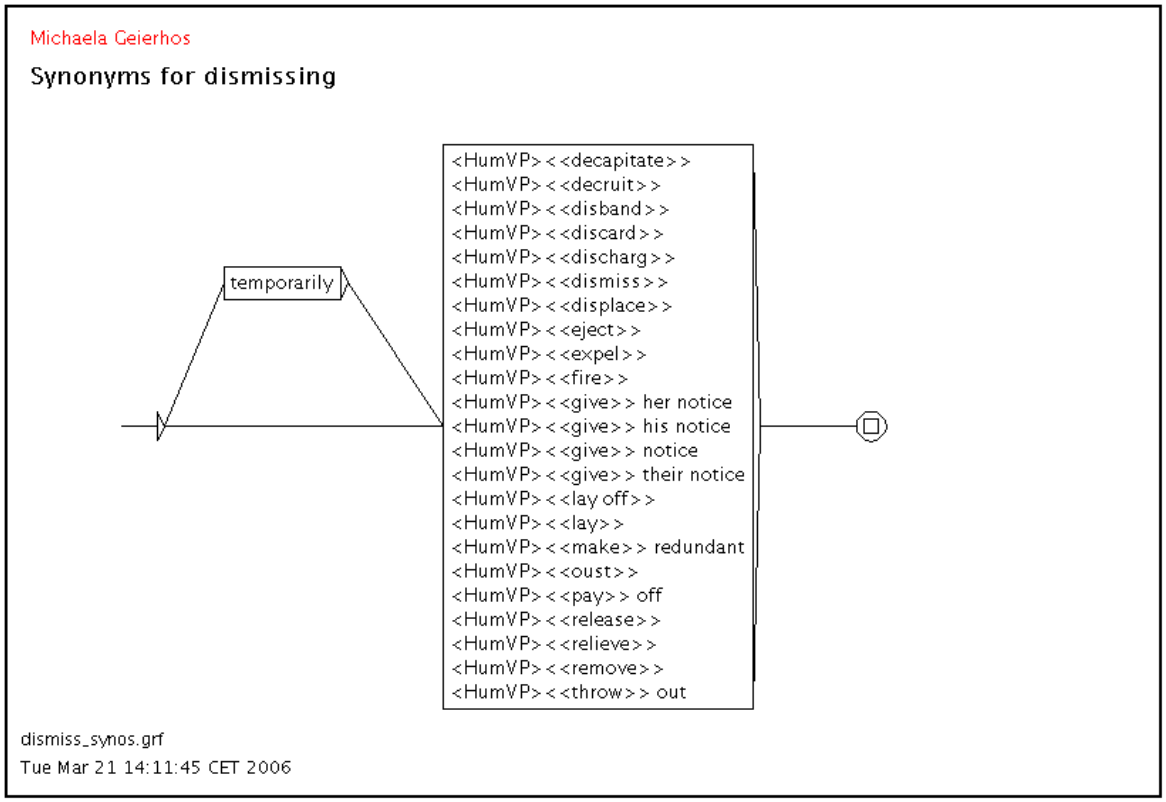


Abbildung 11.12: Graph zur Erkennung von Verben mit der Bedeutung von *„to dismiss s.o.“ – dismiss_synos.grf*

All diese Prädikate wurden in die Grammatik zur Erkennung von „Entlassungsereignissen“ aufgenommen. Allerdings handelt es sich bei diesen Verben um Aktivformen, d.h. dass sie Relationen wiedergeben, in denen eine Firma oder eine Person jemanden aus einem Beschäftigungsverhältnis entlässt. Während der Subgraph *dismiss_synos.grf* aus Abbildung 11.12 die einzelnen Verbkonstruktionen behandelt, analysiert der Hauptgraph *dismiss_as.grf* auf Seite 129 den rechten Kontext dieser Prädikate. Hierbei wird obligatorisch zuerst die zu entlassende Person genannt, bevor möglicherweise Angaben über die Position gemacht werden, welche sie zuvor beansprucht hat. Natürlich lässt sich der gleiche Sachverhalt auch in einer passiven Äußerung darstellen. In diesem Fall nimmt die entlassene Person die Stelle es Subjekts im Satz ein, und die Information, wer sie gekündigt hat, fällt dann meist weg. Im Wesentlichen verhalten sich die endlichen Automaten für diesen Typ Satz ganz analog zu der eben vorgestellten Phrasenstruktur und bedürfen deshalb keiner weiteren Erklärung.

```
The Russian president was expected to dismiss <Person> <PersonName>Mikhail
Kasyanov</PersonName> </Person> as <JD>prime minister</JD> but not until after the
March 14 election.
that the decision to fire <Person> <PersonName>Mikhail Brudno</PersonName> </Person>
as <JD>interim head</JD> of refining and marketing operations was dictated by the
company's aim
yesterday with his surprise decision to fire <Person> <PersonName>Mikhail
Kasyanov</PersonName> </Person> as <JD>prime minister</JD>.
```

```
executive Sir Geoff Mulcahy (left), has fired <Person> <PersonName>Numis</PersonName>
            </Person> as <JD> adviser</JD> on its pounds 50m flotation.
led by Fidelity International, ousted <Person> <PersonName>Michael Green</PersonName>
             </Person> as <JD>chairman-designate</JD> four months ago.
      Disney has stepped up his campaign to remove <Person> <PersonName>Michael
      Eisner</PersonName> </Person> as <JD>chief executive</JD> of Walt Disney
```

11.3.2 Die Nachfolge: „to be replaced as“

In großen Unternehmen ist es durchaus keine Seltenheit, dass von Zeit zu Zeit ein Machtwechsel stattfindet. Dabei liest man oft als Schlagzeile, dass wieder ein Generationenwechsel vollzogen wurde. Doch im Grunde wird hier nur eine Nachfolge für einen Firmenposten geregelt. Immer wenn jemand ein Amt abgibt, egal ob freiwillig oder ungewollt, muss dessen Stelle neu besetzt werden. Diese eben beschriebene Relation wird durch das Prädikat *„to be replaced“* ausgedrückt. Welche Faktoren bei einer derartigen Neu-

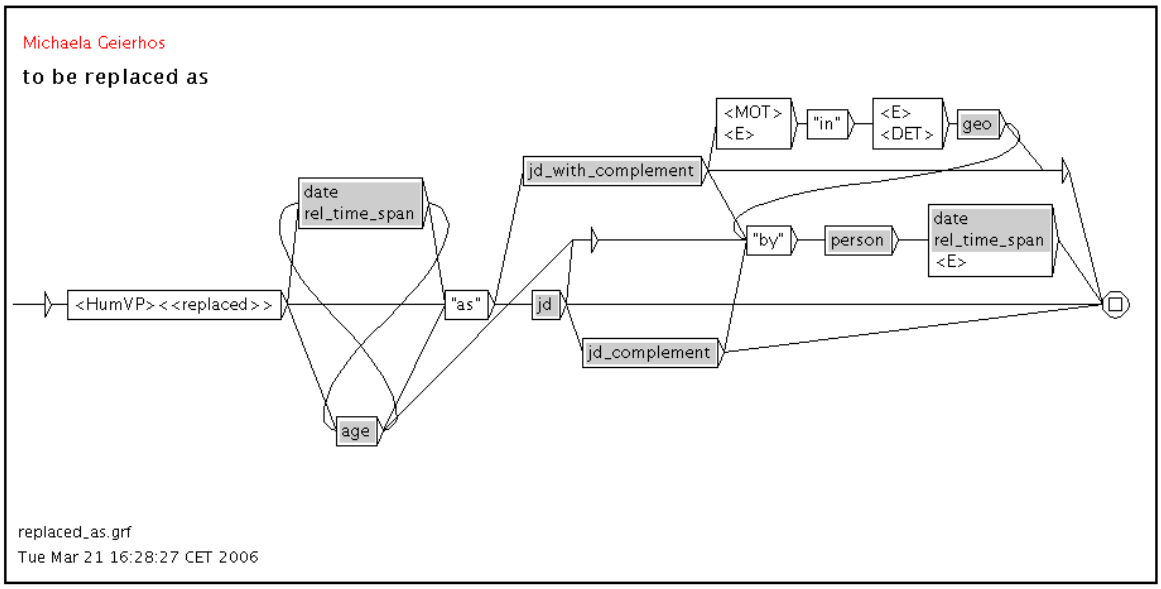


Abbildung 11.13: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to be replaced as“ – replaced_as.grf*

besetzung einer Arbeitsstelle berücksichtigt werden müssen, versucht die Grammatik aus Abbildung 11.13 in den Griff zu bekommen. Denn einerseits sollte erkannt werden, welcher Posten neu zu vergeben ist, was durch den Automaten zur Lokalisierung von Berufsbezeichnungen (siehe Seite 119) erfolgt. Andererseits dürfen Informationen, wie der Zeitpunkt des Wechsels, der Name der betroffenen Firma, und wer der bereits bestimmte Nachfolger ist, auf keinen Fall übergangen werden.

Wie die folgende Konkordanz zeigt, werden all diese Dinge im Kontext des Verbes *to be replaced“* berücksichtigt:

```
KEN BATES has been replaced as <JD>chairman</JD> of <ORG>Chelsea Village</ORG>
   Brian Peace has been replaced as <JD>chief executive</JD> of <ORG>Peace
                                    Software</ORG>
```

He has been replaced as the <JD>chairman</JD> of <ORG>Chelsea Village</ORG> by <Person> <PersonName>Bruce Buck</PersonName> </Person>
Sanderson was replaced as <JD>captain</JD> <Date>last night</Date> by <Person> <PersonName> Pete Anglesea</PersonName> </Person>
M. Le Pen was replaced as <JD>head</JD> of the NF list in the Provence region by Guy Macary
when he was replaced as <JD>President</JD> by <Person> <PersonName>Sir Leonard Hutton</PersonName> </Person> <Date>in 1990</Date>
Mahathir was replaced as <JD>prime minister</JD> <Date>last year</Date> by <Person> <PersonName>Abdullah Ahmad Badawi</PersonName> </Person>
He will be replaced as <JD>chief executive</JD> by <Person><PersonName> Barry Elson</PersonName> </Person>

11.3.3 Die Abdankung: „to resign as“

Wenn jemand darüber nachdenkt, abzudanken, dann ist dies vergleichbar mit einer Entlassung auf eigenen Wunsch. Um dieses Ereignis besser in Worte fassen zu können, wird die dazu entwickelte Grammatik aus Abbildung 11.14 auf der nächsten Seite folgende Prädikate behandeln:

- *to resign*
- *to quit*
- *to leave job*

Zudem werden bei der letztgenannten Verbkonstruktion im Vokabular des Graphen noch weitere Synonym- und Morphemvarianten von *„job“* zugelassen. Darunter fallen die Begriffe *„jobs, post, posts, employment, employments, service, position, positions, office, offices“* und *„work“*.

Auf diese Weise erfasst der Transduktor *resigned.grf* folgende Beispielphrasen:

Rauf Denktas said he had considered resigning as a <JD>negotiator</JD>
his daughter Lea Rose, 23, who had quit her job as a <JD>housemaid</JD> in <GEO>Manila</GEO>
Chancellor Gerhard Schroder has resigned as <JD>chairman</JD> of the <ORG>Social Democratic Party</ORG> <ORG>(SPD)</ORG>
Martin Stewart is to quit as <ORG>BSkyB</ORG>'s <JD>chief financial officer</JD>

11.3.4 Die Pensionierung: „to retire so.“ bzw. „to be retired“

Bei der Pensionierung scheidet zwar eine Person aus Altersgründen aus dem Amt bzw. aus dem Berufsleben aus, doch geschieht dies in der Regel ebenfalls auf freiwilliger Basis.

Um die Tatsache zu beschreiben, dass jemand in Rente geht oder bereits im Ruhestand ist, kommen diese Aktiv- und Passivkonstruktionen folgender englischer Verben in Frage:

- *to retire s.o.*
- *to stop s.o. working*
- *to stop work*
- *to be retired*

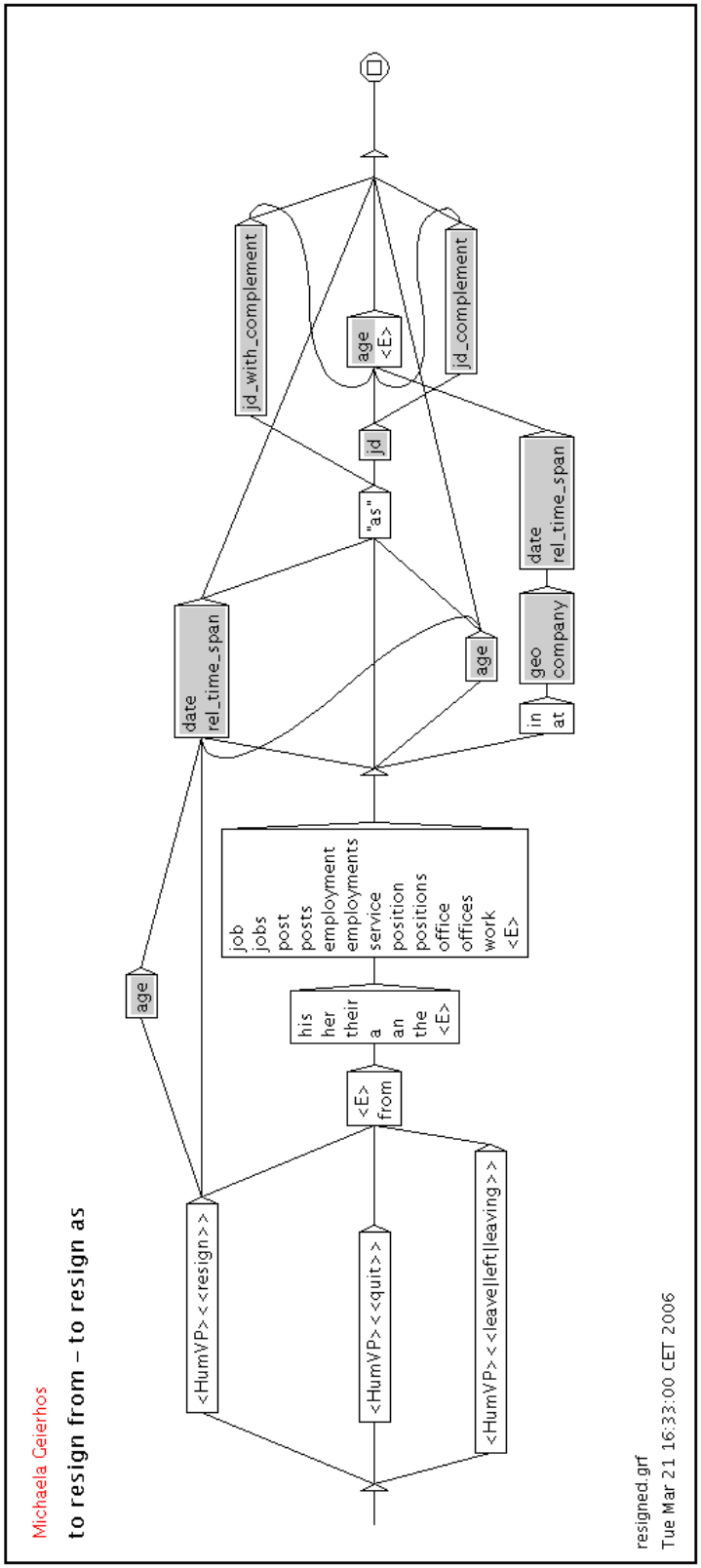


Abbildung 11.14: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to resign (as/from)“* und seiner direkten Synonyme – *resigned.grf*

- *to give up work*
- *to stop work*
- *to be stopped working*
- *to reach retirement age*

Die aktiven Verbalphrasen werden nun von der Grammatik *retire_as.grf* aus Abbildung 11.15 und die passiven Konstruktionen vom Graphen *retired_as.grf* (ohne Abbildung) erfasst. Da sich die beiden Automaten natürlich sehr stark gleichen, können sie hier zusammen behandelt werden, denn jeder von ihnen berücksichtigt das Rentenalter, eventuelle Angaben zum Beruf der jeweiligen Person, wann jemand in Rente gegangen ist, oder wie lange die Pensionierung schon zurückliegt.

Die markierten Verbkonstruktionen mitsamt ihres rechten Kontextes könnten beispielsweise wie folgt aussehen:

write retired <Person> <JD>actuary</JD> <PersonName>Brian FitzGerald</PersonName> </Person> and <Person> <JD>author</JD> <PersonName>Bruce Cohen</PersonName> </Person> including retired <Person> <JD>Assistant Commissioner</JD> <PersonName>Hugh Sreenan</PersonName> </Person>

Minister Yoshiro Hayashi, who retired as a <JD> awmaker</JD> <Date>last year</Date>

Denis Hurley retired as <JD>Archbishop</JD> <Date>in 1992</Date>

Retired Kenyan <Person> <JD>athlete</JD> <PersonName>Rose Tata Muya</PersonName> </Person>

Rocky Marciano, have retired as <JD>champion</JD> and stayed retired

Luzviminda Tancangco who retired <Date>on Feb. 2</Date>

Denis Brosnan retired as <JD>Chairman</JD> and <JD>Director</JD> of the Group

David Creary, who retired as <JD>chief operating officer</JD> <Date>last week</Date>

The 17-year-old son of retired Croatian <Person> General <PersonName>Vladimir Zagorac</PersonName> </Person> was kid

Minister Dan Naveh has appointed retired <Person> <ORG>Jerusalem District Court</ORG> <JD>president</JD> <PersonName>Vardi Zeiler</PersonName> </Person>

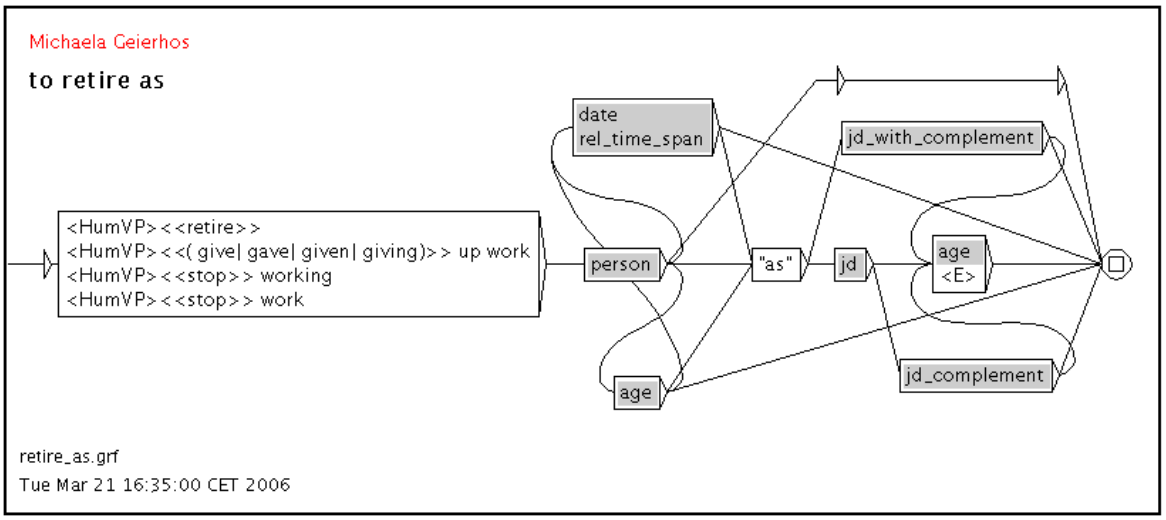


Abbildung 11.15: Graph zur Erkennung von Verbalphrasen mit dem Verb *„to retire s.o.“* und seiner direkten Synonyme – *retire_as.grf*

12 Auswertung der Ergebnisse

12.1 Evaluationsmaße [Wikipedia, 2005/2006]

Vollständigkeit und Genauigkeit sind zwei Maße zur Beschreibung der Güte eines Suchergebnisses beim Information-Retrieval oder bei einer Recherche im Allgemeinen.

Für die Evaluierung eines Information-Retrieval-Systems sollten die beiden zusammenhängenden Maße gemeinsam betrachtet werden. In der Regel sinkt mit steigendem Recall (mehr Treffer) die Precision (mehr irrelevante Ergebnisse) und umgekehrt sinkt mit steigender Precision (weniger irrelevante Ergebnisse) der Recall (mehr relevante Dokumente, die nicht gefunden werden). Somit besteht eine negative Korrelation.

Stellt man das Verhältnis zwischen Recall und Precision in einem Diagramm dar, so wird der (höchste) Wert im Diagramm, an dem der Precision-Wert gleich dem Recall-Wert ist - also der Schnittpunkt des Precision-Recall-Diagramms mit der Identitätsfunktion - der Precision-Recall-Breakeven-Punkt genannt.

Für die Evaluierung des Information-Retrieval-Systems gibt es mit dem Fall-Out noch ein drittes Kriterium.

12.1.1 Precision bzw. Genauigkeit

Die Precision beschreibt die Genauigkeit eines Suchergebnisses. Sie ist definiert als der Anteil der gefundenen relevanten Dokumente von allen bei einer Suche gefundenen Dokumenten.

$$\text{Precision} = \frac{|\{\text{relevante Dokumente}\} \cap \{\text{gefundene Dokumente}\}|}{|\{\text{gefundene Dokumente}\}|}$$

12.1.2 Recall bzw. Vollständigkeit

Der Recall beschreibt die Vollständigkeit eines Suchergebnisses. Er ist definiert als der Anteil bei einer Suche gefundenen relevanten Dokumente bzw. Datensätze an den relevanten Dokumenten der Grundgesamtheit.

$$\text{Recall} = \frac{|\{\text{relevante Dokumente}\} \cap \{\text{gefundene Dokumente}\}|}{|\{\text{relevante Dokumente}\}|}$$

12.1.3 Fall-Out

Das Fall-Out beschreibt in negativer Weise die Güte des zu bewertenden Verfahrens, indem die Anzahl der gefundenen irrelevanten Dokumente durch die Gesamtanzahl ir-

relevanter Dokumente geteilt wird.

$$\text{Fall-Out} = \frac{|\{\text{irrelevante Dokumente}\} \cap \{\text{gefundene Dokumente}\}|}{|\{\text{irrelevante Dokumente}\}|}$$

12.2 Qualität des Systems

Um ein Verständnis dafür zu erlangen, wie präzise die im Laufe dieser Arbeit vorgestellten Automaten auf dem FT-Korpus arbeiten, müssen die Suchergebnisse entsprechend ausgewertet werden.

Für die nun folgende Beispielauswertung der Treffermenge des Automaten *person_name.grf* aus Abbildung 6.1 (siehe Seite 71) wird mithilfe der Evaluationsmaße Precision und Recall die Qualität dieser Grammatik in Bezug auf das Korpus ermittelt.

Hierfür ist es notwendig, die tatsächliche Anzahl an Vorkommen einer gesuchten Entität im Text zu kennen. Deshalb ist es kaum möglich diese Auswertung auf dem kompletten Financial Times Korpus durchzuführen, und es ist ratsam, sich nur auf einen Teil des Korpus zu konzentrieren. Dieses Teilkorpus sollte nicht verstärkt zum Training der Automaten eingesetzt worden sein, da sonst kein repräsentatives Auswertungsergebnis erzielt werden könnte. Denn würde die Evaluation auf dem gleichen Text erfolgen, welcher schon während der Entwicklung der Grammatiken für Zwischentests eingesetzt wurde, wären herausragende Werte sicherlich eine direkte Konsequenz bei diesem Vorgehen.

Auch lässt sich plausibel erklären, warum diese Auswertung ausgerechnet mit dem endlichen Automaten zur Erkennung von Personennamen erfolgen soll. Wie bereits in Kapitel 6 diskutiert wurde, ist der Einsatz eines Automaten ohne größeres Kontextwissen ein sehr risikobehaftetes Unterfangen. Da der Personennamengraph viel weniger Kontextinformation als der Personengraph einsetzt und ohne die Kontexte der Verbalphrasenautomaten aufgerufen wird, müsste er verhältnismäßig schlechtere Ergebnisse erzielen.

Die folgende statistische Auswertung dieser Treffermenge soll nun folgendes beweisen: Wenn eine Grammatik auf qualitativ gute Lexikoneinträge setzen kann, fällt die vernachlässigte Kontextinformation nicht so stark in das Gewicht der Auswertung. Denn wäre die Wörterbuchbasis, welche der Automat nutzen konnte, wesentlich kleiner gewesen, könnten sich die Werte für Precision und Recall nicht mehr sehen lassen.

Tatsächliche Anzahl der Personennamen im Text	4267
Anzahl der gefundenen Personennamen im Text	4561
Anzahl der korrekt gefundenen Personennamen	3916
Recall	85,85%
Precision	91,77%

Tabelle 12.1: Statistische Auswertung der Extraktionsresultate

13 Anwendungen

Nachdem nun eine Reihe an verschiedenen ausgewählten Verbrelationen präsentiert wurde, ist es an der Zeit weitere Verbkonstruktionen zu ermitteln, welche mit den hier behandelten Entitäten wie z.B. den Personen und Organisationen einhergehen.

Die folgenden Graphen versuchen mithilfe einer groben Schematisierung des potentiellen Kontextes von personenbezogenen Prädikaten, diese auf relativ einfache und schnelle Weise automatisch aus den Korpora zu extrahieren.

13.1 Automatische Extraktion von Relationen zwischen Personen und Organisationen

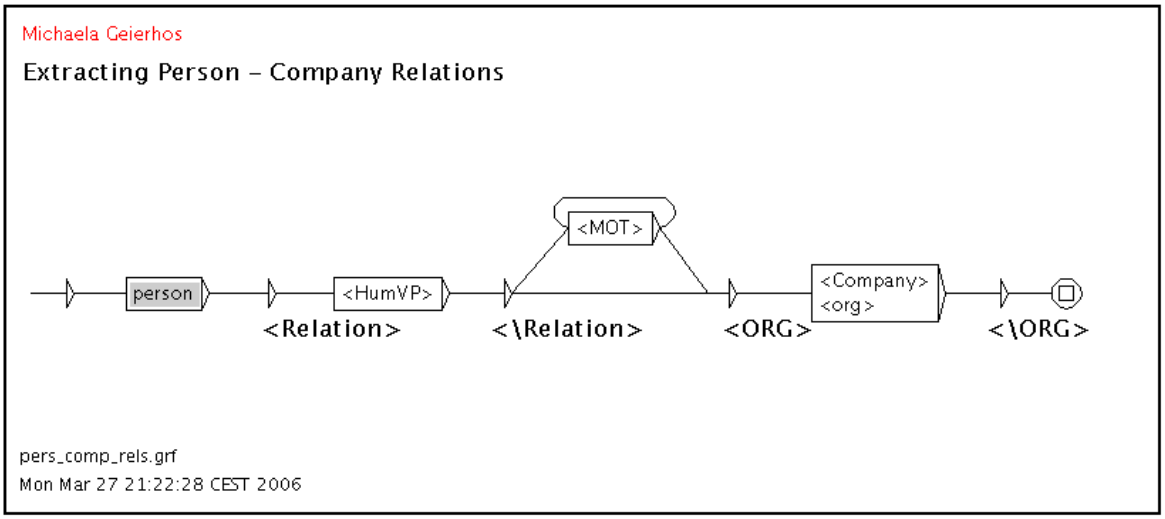


Abbildung 13.1: Graph zur automatischen Erkennung weiterer Verbkonstruktionen, die eine Beziehung zwischen einer Person und einer Firma ausdrücken

Im Laufe meiner Untersuchungen von biographischen Relationen stellte sich heraus, dass es ratsam ist, das Suchfenster nicht zu weit zu fassen, gerade wenn neue Informationen zwischen zwei Entitäten gewonnen werden sollen.

Zwar scheint der Graph aus Abbildung 13.1 auf den ersten Blick recht einfach aufgebaut zu sein, doch liefert er sehr gute Treffer bei der Erkennung potentieller Verbrelationen, welche im Zusammenhang mit Menschenbezeichnern und Firmennamen stehen.

Um die Effizienz des Automaten bei der Suche nach Kandidaten für Personen-Firmen-Relationen nicht allzu sehr zu beeinträchtigen, wurde auf den Einsatz des *company.grf*-Graphen verzichtet. Da es hierbei nicht um die genaue Lokalisierung von Organisations-

namen geht, reicht es mit den Symbolen <Company> und <org> aus den Lexika zu arbeiten und vorangehende Textpassagen mit dem „Platzhalter“ <MOT>* zu versehen.

Auf diese Weise lassen sich folgende Sequenzen im Korpus aufspüren und aufgrund ihrer Annotation automatisch weiterverarbeiten:

<Person> <PersonName>Brown<\PersonName> <\Person> <Relation>has followed<\Relation> <ORG> Leeds United<\ORG>

<Person> <PersonName>Canon Andrew White<\PersonName> <\Person> <Relation>said<\Relation> in an interview with British <ORG>Broadcasting Corp.<\ORG>

<Person> <PersonName>Gianna Angelopoulos-Daskalaki<\PersonName> <\Person> <Relation>told<\Relation> <ORG> The Associated Press<\ORG>

<Person> <PersonName>Greg Dyke<\PersonName> <\Person> <Relation>has transformed<\Relation> the <ORG> BBC<\ORG>

13.2 Automatische Extraktion von Relationen zwischen mindestens zwei Personen

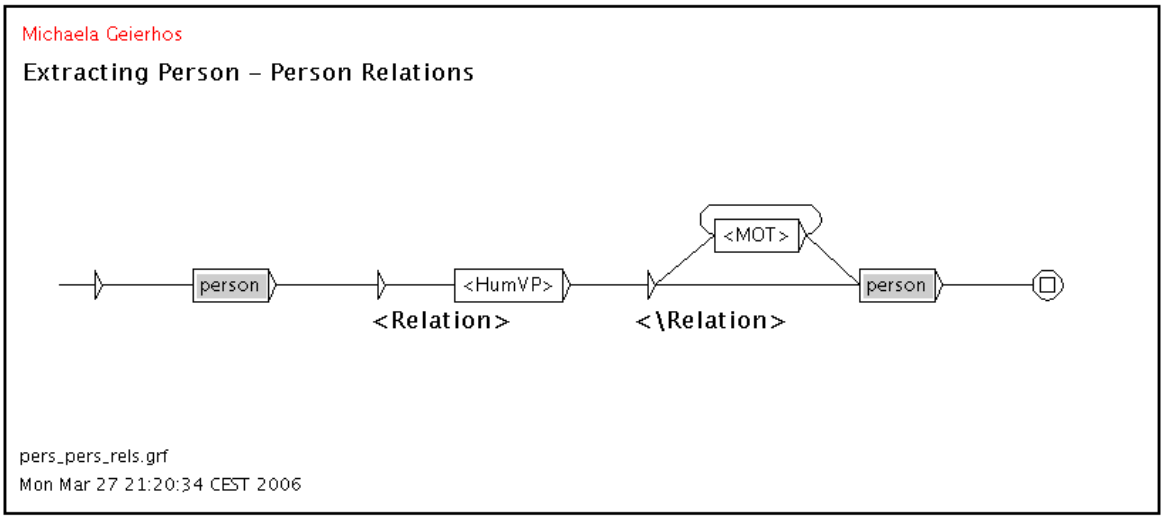


Abbildung 13.2: Graph zur automatischen Erkennung weiterer Verbkonstruktionen, die eine Beziehung zwischen mindestens zwei Personen ausdrücken

Das Gleiche gilt auch für die Beziehung zwischen zwei Menschen. Jedes Verb, welches an beiden Argumentpositionen einen Menschenbezeichner verlangt, sollte in einer Konkordanz dieses Graphen enthalten sein. Jedoch stellt dieser Automat - so wie schon sein Vorgänger - die Bedingung, dass die potentiellen Verben aus dem Lexikon der personenbezogenen Prädikate stammen müssen. Das ist unter anderem ein Grund dafür, warum tatsächlich nur Verben erkannt werden.

Leider hat diese Extraktionsmethode einen kleinen „Schönheitsfehler“, da sie einen Zyklus in einer freien Variable wie <MOT> zulässt. So kann es passieren, dass der Kontext zu stark ausgedehnt wird und der Bezug zum jeweiligen Verb verloren geht. Wenn dies jedoch nicht der Fall ist, bekommt man ausgezeichnete Ergebnisse wie diese:

<Person>runner-up <PersonName>Charlie Maher<\PersonName> <\Person> <Relation>will be accompanied<\Relation> by a canine companion at <Person> <PersonName>Ron Chereskin<\PersonName> <\Person>

<Person> <PersonName>Abu Salem<\PersonName> <\Person> <Relation>has been declared<\Relation> a terrorist by a <Person>Portuguese<\Person>

13.3 Automatische Extraktion von Relationen zwischen Personen und ihren Berufen

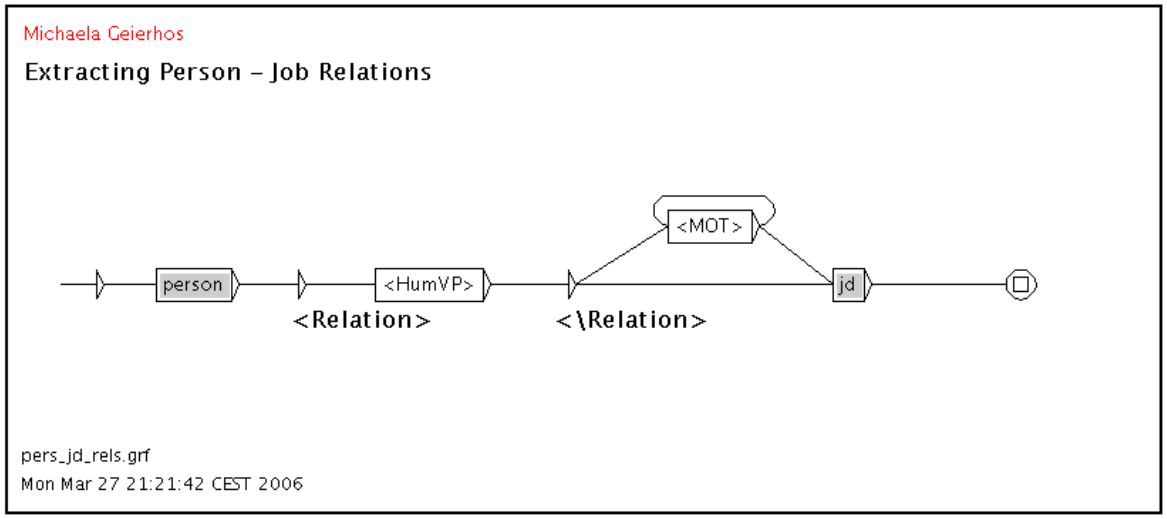


Abbildung 13.3: Graph zur automatischen Erkennung weiterer Verbkonstruktionen, die eine Beziehung zwischen einer Personen und ihrem Beruf ausdrücken

Nachdem hier die Thematik der Berufsbezeichnungen und den dazugehörigen Arbeitsverhältnissen ausgiebig behandelt wurde, bietet es sich noch an, einen weiteren Transduktor zu erstellen, welcher sich auf das Finden neuer Verbrelationen zwischen Menschenbezeichnern und ihren jeweiligen beruflichen Tätigkeiten spezialisiert.

So können damit beispielsweise folgende Treffer erzielt werden:

<Person><PersonName>Yigal Berman<\PersonName> <\Person> <Relation>has provided<\Relation> to the Company during his tenure as a <JD>director<\JD>

<Person> <PersonName>Ted Kunkel<\PersonName> <\Person> <Relation>has been<\Relation> an <JD> outstanding leader<\JD>

14 Zusammenfassung und Ausblick

Das Ziel dieser Arbeit war es Menschenbezeichner innerhalb biographischer Relationen in englischsprachigen Wirtschaftsnachrichten automatisch zu erkennen.

In diesem Zusammenhang wurde der Begriff der „biographischen Relation“ ausführlich diskutiert, wobei dessen einzelne Facetten zum Vorschein kamen. Hierbei wurde ersichtlich, dass ein biographischer Kontext recht vielfältig sein kann und die Grammatikentwicklung zunächst ihre Tücken hat. Doch letztendlich besteht eine Möglichkeit, das textuelle Umfeld von Personen in einige Hauptkategorien zu unterteilen, und auf ihnen basierend erste linguistische Untersuchungen vorzunehmen.

Auf diese Weise können wiederum einzelne Verbkonstruktionen innerhalb dieser Gruppen, syntaktisch oder semantisch zusammengefasst, bzw. Ähnlichkeiten unter ihnen ausgemacht werden. Genauso verlief auch die Arbeit mit der Grammatik der beruflichen Relationen. Nach diversen Vorüberlegungen, welche Prädikate in die jeweilige Kategorie fallen, mussten diese semantisch klassifiziert, und im Anschluss an die strukturelle Analyse schematisiert und modularisiert werden. So konnten sich Teilgrammatiken in den verschiedensten Automaten wiederfinden, und wurden anhand diverser Kontexte konzeptionell und inhaltlich verbessert.

Für die hier ausgewählten Relationen ließ sich das Problem der automatischen Lokalisierung von Personen mit biographischen Kontexten relativ gut bewältigen. Außerdem liefern die entsprechenden Grammatiken zufriedenstellende Ergebnisse für die einzelnen Korpora.

Jedoch sollte man sich dessen bewusst sein, dass mit dieser Arbeit nur die „Spitze des Eisbergs“ angekratzt wurde. Auch wenn das Feld der biographischen Relationen überschaubar ist, weil es eine endliche, nicht allzu große Menge von ihnen gibt, ist es ihr Kontext, welcher erst einmal analysiert und in Form von lokalen Grammatiken formalisiert werden muss. Je mehr man über diese Relationen im einzelnen auf syntaktischer und semantischer Ebene herausfindet, desto besser werden die Ergebnisse von „intelligenten“ Informationsextraktionssystemen.

Mit dieser Arbeit wurde der erste Grundstein für sich weiterentwickelnde und umfassendere Grammatiken zur Erkennung von Menschenbezeichnern und ihrer biographischen Verbrelationen gelegt. Wobei sich das hier vorgestellte Konzept ohne größere Schwierigkeiten auf andere personenbezogene Prädikate, wie z.B. *„to stay with so., to live, to restore so. to life, to kill so., to laugh, to be moved to tears, to read sth., to cook sth., to sing sth., to kiss so., to love so., to visit so., to feat sth., to repare sth. “* übertragen lässt. Natürlich gibt es noch eine Reihe von Prädikaten, welche aus biographischer Sicht noch viel relevanter sind und ebenfalls untersucht werden sollten. Diese Verbkonstrukte können zukünftig Gegenstand weiterer linguistischer Studien sein und müssen hier nicht mehr aufgeführt werden.

A Übersicht aller Kategorien in den Wörterbüchern

A.1 Semantische Kategorien

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
Abbrev	semantisch	Abbreviation	Abkürzung
ABourough	semantisch	Bourough Adjective	Adjektiv für einen Stadtteil
ACity	semantisch	City Adjective	Adjektiv für eine Metropole, Stadt
AGEO	semantisch	Geographical Term Adjective	Adjektiv für ein Toponym, einen geographischen Begriff
ANation	semantisch	Nation Adjective	Adjektiv für ein Land
AProvince	semantisch	Province Adjective	Adjektiv für eine Provinz
AState	semantisch	State Adjective	Adjektiv für einen Bundesstaat
AuProvinceCitizen	semantisch	Australian Province Citizen	Bewohner einer australischen Provinz
Borough	semantisch	Borough	Stadtteil, Stadtbezirk
CaProvinceCitizen	semantisch	Canadian Province Citizen	Bewohner einer kanadischen Provinz
CaProvince	semantisch	Canadian Province	Kanadische Provinz
Citizen	semantisch	Citizen	(Staats)Bürger
City	semantisch	City	Metropole, Stadt
Company	semantisch	Company Name	Organisationsname bzw. Firmenname
Continent	semantisch	Continent	Kontinent
County	semantisch	County	Grafschaft
Discipline	semantisch	Discipline	Fachbereich, Lehrbereich
Département	semantisch	Département	Département
FN	semantisch	First Name	Vorname
GEO	semantisch	Geographical Term	Toponym, geographischer Begriff
Hum	semantisch	Human	Menschenbezeichner
JD	semantisch	Job Descriptor	Berufsbezeichner

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
LN	semantisch	Long Name	Vollständiger Personenname, bestehend aus Vor- und Nachname, evtl. mit Titel
Nation	semantisch	Nation	Land
NYCBourough	semantisch	New York City Bourough	Stadtteil von New York City
NYCcitizen	semantisch	New York City Citizen	Bewohner von New York City
PR	semantisch	Proper Noun	Eigenname
Region	semantisch	Region	Region, Gebiet
Sector	semantisch	Sector	Sektor, Branche
SN	semantisch	Surname	Nachname
Title	semantisch	Title	Titel, wie z.B. akademische Grade, aristokratische Titel
Urbanite	semantisch	Urbanite	Stadtbewohner
USstate	semantisch	US State	US Bundesstaat
USstateCitizen	semantisch	US State Citizen	Bewohner eines US Bundesstaates

A.2 Grammatikalische Kategorien

Abkürzung	Kategorietyp	wörtliche Bedeutung	Erläuterung
A	grammatikalisch	Adjective	Adjektiv
N	grammatikalisch	Noun	Nomen
XA	grammatikalisch	Extended Adjective	lexikalische Einheit, welche die Funktion eines Adjektivs erfüllt
XN	grammatikalisch	Extended Noun	Mehrwortlexem

B Syntaktische Variabilität am Beispiel von „Bill Gates“

1. B. Gates
2. B. Gates III
3. B. Gates, III
4. B. Gates III, KBE
5. B. Gates, KBE
6. B. H. Gates
7. B. H. Gates III
8. B. H. Gates, III
9. B. H. Gates III, KBE
10. B. H. Gates, KBE
11. Bill Gates
12. Bill Gates III
13. Bill Gates, III
14. Bill Gates III, KBE
15. Bill Gates, KBE[42]
16. Bill Henry Gates
17. Bill Henry Gates III
18. Bill Henry Gates, III
19. Bill Henry Gates III , KBE
20. Bill Henry Gates, KBE
21. Bill H. Gates
22. Bill H. Gates III
23. Bill H. Gates, III
24. Bill H. Gates III, KBE
25. Bill H. Gates , KBE
26. Bill (William Henry) Gates
27. Bill (William Henry) Gates III
28. Bill (William Henry) Gates, III
29. Bill (William Henry) Gates III, KBE
30. Bill (William Henry) Gates, KBE
31. Bill (William H.) Gates
32. Bill (William H.) Gates III
33. Bill (William H.) Gates, III
34. Bill (William H.) Gates III, KBE
35. Bill (William H.) Gates, KBE
36. Gates, Bill (William H.)
37. Gates, Bill (William Henry)
38. Gates, Bill (William Henry) III
39. Gates, Bill (William Henry), III
40. Gates, Bill (William H.) III
41. Gates, Bill (William H.), III
42. Gates, William H.

[42] *Knights Commander of the British Empire*

43. Gates, William Henry
44. Gates, William Henry III
45. W. H. Gates
46. W. H. Gates III
47. W. H. Gates, III
48. W. H. Gates III, KBE
49. W. H. Gates, KBE
50. William (Bill) Henry Gates
51. William (Bill) Henry Gates III
52. William (Bill) Henry Gates, III
53. William (Bill) Henry Gates III, KBE
54. William (Bill) Henry Gates, KBE
55. William (Bill) H Gates
56. William (Bill) H. Gates
57. William (Bill) H Gates III
58. William (Bill) H Gates, III
59. William (Bill) H. Gates III
60. William (Bill) H. Gates, III
61. William (Bill) H Gates III, KBE
62. William (Bill) H. Gates III, KBE
63. William (Bill) H Gates, KBE
64. William (Bill) H. Gates, KBE
65. William Gates
66. William Gates III
67. William Gates, III
68. William Gates III, KBE[43]
69. William Gates, KBE
70. William „Bill“ Henry Gates
71. William „Bill“ Henry Gates III
72. William „Bill“ Henry Gates, III
73. William „Bill“ Henry Gates III, KBE
74. William „Bill“ Henry Gates, KBE
75. William „Bill“ H Gates
76. William „Bill“ H. Gates
77. William „Bill“ H Gates III
78. William „Bill“ H Gates, III
79. William „Bill“ H. Gates III
80. William „Bill“ H. Gates, III
81. William „Bill“ H Gates III, KBE
82. William „Bill“ H. Gates III, KBE
83. William „Bill“ H Gates, KBE
84. William „Bill“ H. Gates, KBE
85. William Henry Gates
86. William Henry Gates III
87. William Henry Gates, III
88. William Henry Gates III, KBE
89. William Henry Gates, KBE
90. William H. Gates
91. William H. Gates III
92. William H. Gates, III
93. William H. Gates III, KBE
94. ...

[43] vgl. `http://seattletimes.nwsource.com/html/editorialsopinion/2001845028_billed28.html`

Literaturverzeichnis

[Agichtein und Gravano, 2000] Eugene Agichtein und Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, pages 85–94, 2000.

[Alani *et al.*, 2003] H. Alani, S. Kim, D. E. Millard, M. J. Weal, P. H. Lewis, W. Hall und N. R. Shadbolt. Automatic extraction of knowledge from web documents. In *Proceedings of 2nd International Semantic Web Conference - Workshop on Human Language Technology for the Semantic Web abd Web Services*, Sanibel Island, Florida, USA, 2003.

[Barzilay, 2003] Regina Barzilay. *Information Fusion for Multidocument Summarization: Paraphrasing und Generation.* Dissertation, Columbia University, 2003.

[Biography.com, 2005/2006] Biography.com, 2005/2006. `http://www.biography.com`.

[Blanc und Dister, 2004] Olivier Blanc und Anne Dister. Automates lexicaux avec structure de traits. In *RECITAL 2004*, Fès, 21 avril 2004.

[Blank, 1997] Ingeborg Blank. *Computerlinguistische Analyse mehrsprachiger Fachtexte, CIS-Bericht-98-109.* CIS-Bericht-98-109, Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians-Universität München, 1997.

[Callan, 2005] Jamie Callan. Human language technologies: Information extraction. Lecture Notes, Fall 2005.

[CareerBuilder.com, 2005] CareerBuilder.com, 2005. `http://www.careerbuilder.com`.

[Ciaramita und Altun, 2005] Massimiliano Ciaramita und Yasemin Altun. Named-Entity Recognition in Novel Domains with External Lexical Knowledge. In *Workshop on Advances in Structured Learning for Text und Speech Processing.* NIPS, 2005.

[Constant, 2000] Matthieu Constant. Description d'expressions numériques en français. In Anne Dister, Hrsg., *Revue Informatique et Statistique dans les Sciences humaines 36, Actes des troisièmes journées INTEX*, pages 119–135, Liège, 2000.

[Constant, 2003] Matthieu Constant. *Grammaires locales pour l'analyse automatique de textes : Méthodes de construction et outils de gestion.* Dissertation, Université de Marne la Vallée, 2003.

[Constant, 2004] Matthieu Constant. GRAAL, une bibliothèque de graphes : mode d'emploi. In C. Muller, J. Royeauté und M. Silberztein, Hrsg., *Cahiers de la MSH*

Ledoux 1, INTEX pour la linguistique et le traitement automatique des langues, pages 321–330, Besançon, 2004. Presse Universitaire de Franche-Comté.

[Courtois, 2004] Blandine Courtois. Dictionnaires électroniques DELAF anglais et français. In Christian Leclère, Éric Laporte, Mireille Piot und Max Silberztein, Hrsg., *Lexique, syntaxe et lexique-grammaire; syntax, lexis & lexicon-grammar*, pages 113–123. John Benjamins Publishing Company, 2004.

[Cruse, 1986] A. D. Cruse. *Lexical Semantics*. Cambridge University Press, 1986.

[Cucchiarelli und Velardi, 2001] A. Cucchiarelli und P. Velardi. Unsupervised named entity recognition using syntactic und semantic contextual evidence. *Computational Linguistics*, Volume 27(Number 1), 2001.

[Cucerzan und Yarowsky, 1999] S. Cucerzan und D. Yarowsky. Language independent named entity recognition combining morphological und contextual evidence. In *Proceedings of the Joint SIGDAT Conference on EMNLP und VLC*, pages 90–99, 1999.

[Danlos und Gross, 1988] Laurence Danlos und Maurice Gross. Building Electronic Dictionaries for Natural Language Processing. In L. Kott, Hrsg., *Proceedings of the 2nd Symposium France-Japan*. Amsterdam: North Holland, 1988.

[DoPL, 2005] DoPL. Division of Professional Licensure - List of Professions. Online-Verzeichnis, 2005. `http://www.mass.gov/dpl/boards/dirprofs.htm`.

[DOT, 2005] DOT. Dictionary Of Occupational Titles. Online-Verzeichnis, 2005. `http://www.occupationalinfo.org/`.

[Drosdowski *et al.*, 2006] Matthias Wermke, Kathrin Kunkel-Razum und Wolfgang Müller, Hrsg. *DUDEN - Rechtschreibung der deutschen Sprache*. Dudenverlag, Mannheim, 2006. 24. Auflage. Neue Rechtschreibung.

[Duboué und McKeown, 2003] Pablo Duboué und Kathleen McKeown. Statistical Acquisition of Content Selection Rules for Natural Language Generation. In *Proceedings of the 2003 Conference on Empirical Methods for Natural Language Processing (EMNLP 2003)*, Sapporo, Japan, July 2003.

[Duboué *et al.*, 2003] Pablo Duboué, Kathleen McKeown und Vasileios Hatzivassiloglou. ProGenIE: Biographical descriptions for intelligence analysis. In *Proceedings of the NSF/NIJ Symposium on Intelligence und Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 343–345, Tucson, Arizona, USA, June 2003. Springer-Verlag.

[Fairon, 2000] Cédrick Fairon. *Structures non-connexes. Grammaire des incises en français : description linguistique et outils informatiques.* Dissertation, Université Paris 7, 2000.

[Fellbaum, 1998] Christiane Fellbaum, Hrsg. *WordNet - An Electronic Lexical Database*. MIT Press, 1998.

[Friburger und Maurel, 2001] Nathalie Friburger und Denis Maurel. Elaboration d'une cascade de transducteurs pour l'extraction des noms personnes dans les textes. In *TALN 2001*, Tours, 2-5 juillet 2001.

[Friburger, 2002] Nathalie Friburger. *Reconnaissance automatique des nomes propres - Application à la classification automatique de textes journalistiques.* Dissertation, Université François Rabelais, Tours, 2002.

[Gaizauskas, 2002] Robert Gaizauskas. Information extraction an information extraction perspective on text mining: Tasks, technologies und prototype applications. Euromap Text Mining Seminar, September 4, 2002.

[Geierhos, 2004] Michaela Geierhos. Einführung in WordNet 2.0. `http://www.cis.uni-muenchen.de/~micha/old/WordNet.pdf`, 3. Februar 2004.

[Geierhos, 2005] Michaela Geierhos. DELA Wörterbücher: Der Umgang mit externen Ressourcen in Unitex. Was man beim Erstellen eigener Lexika beachten sollte? `http://www.cis.uni-muenchen.de/~micha/old/Dela.pdf`, 2. Mai 2005.

[Government of Newfoundland und Labrador, 2005] Government of Newfoundland und Labrador. Canadian job classification. Online-Verzeichnis, 2005. `http://www.fin.gov.nl.ca/fin/pensions/pdf/employer_specs/job_class.pdf`.

[Gross und Senellart, 1998] Maurice Gross und Jean Senellart. Nouvelles bases pour une approche statistique. In *JADT98*, Nice, France, 1998.

[Gross, 1975] Maurice Gross. On the relations between syntax und semantics. In E. L. Keenan, Hrsg., *Formal Semantics of Natural Language*, pages 389–405. Cambridge: Cambridge University Press, 1975.

[Gross, 1977] Maurice Gross. Remarks on the separation between syntax und semantics. In *Studies in Descriptive und Historical Linguistics, Festschrift for Winfred P. Lehmann*, pages 71–81, Amsterdam/Philadelphia, 1977. Benjamins.

[Gross, 1978] Maurice Gross. Taxonomy in syntax. *SMIL, Journal of Linguistic Calculus 1978:3-4*, pages 73–96, 1978. Stockholm: Skriptor.

[Gross, 1979] Maurice Gross. On the failure of generative grammar. *Language 55:4*, pages 859–885, 1979.

[Gross, 1983] Maurice Gross. On structuring the lexicon. *Quaderni di Semantica 4:1*, pages 107–120, 1983.

[Gross, 1986] Maurice Gross. Lexicon-Grammar: The Representation of Compound Words. In *COLING-1986 Proceedings*, pages 1–6, 1986.

[Gross, 1988] Maurice Gross. Methods und Tactics in the Construction of a Lexicon-Grammar. *Linguistics in the Morning Calm 2, Selected papers from SICOL 1986*, pages 177–197, 1988. Seoul: Hanshin.

[Gross, 1991] Maurice Gross. Linguistic representations und text analysis. *Linguistic Unity und Linguistic Diversity in Europe*, pages 31–61, 1991. London: Academia Europaea.

[Gross, 1992] Maurice Gross. The argument structure of elementary sentences. *Language Research 28:4*, pages 699–716, 1992. Seoul National University.

[Gross, 1993] Maurice Gross. Local grammars und their representation by finite automata. In M. Hoey, Hrsg., *Data, Description, Discourse, Papers on the English Language in honour of John McH Sinclair*, pages 26–38. Harper-Collins, London, 1993.

[Gross, 1994] Maurice Gross. Constructing lexicon-grammars. In Atkins und Zampolli, Hrsg., *Computational Approaches to the Lexicon*, pages 213–263. Oxford Univ. Press, 1994.

[Gross, 1996] Maurice Gross. Lexicon grammar. In K. Brown und J. Miller, Hrsg., *Concise encyclopedia of syntactic theories*, pages 244–258. Pergamon, 1996.

[Gross, 1997] Maurice Gross. The Construction of Local Grammars. In E. Roche und Y. Schabès, Hrsg., *Finite-State Language Processing*, pages 329–354. Language, Speech und Communication, Cambridge, Mass.: MIT Press, 1997.

[Gross, 1998–1999] Maurice Gross. Lemmatization of compound tenses in English. *Lingvisticae Investigationes*, XXII:71–122, 1998-1999.

[Gross, 1999] Maurice Gross. A bootstrap method for constructing local grammars. In *Contemporary Mathematics: Proceedings of the Symposium, University of Belgrad*, pages 229–250, Belgrad, 1999.

[Guenthner, 1996] Franz Guenthner. Electronic lexica und corpora research at cis. *International Journal of Corpus Linguistics*, 1(2):287–301, 1996.

[Guide to the World of Occupations, 2005] Guide to the World of Occupations. Alphabetical list of occupations, 2005. `http://www.occupationsguide.cz/en/abecedni/abecedni.htm`.

[Hopcroft *et al.*, 2002] John E. Hopcroft, Rajeev Motwani und Jeffrey D. Ullman. *Einführung in die Automatentheorie, Formale Sprachen und Komplexitätstheorie*. Addison-Wesley Pearson Studium, 2002.

[LabourMarket, 2005] LabourMarket. Online-Verzeichnis, 2005. `http://www.labourmarket.co.nz/labourmarket.htm`.

[Lacombe, 2004] J. Lacombe. List of occupations. Online-Verzeichnis, 2004. `http://www.cpcug.org/user/jlacombe/terms.html`.

[Langer, 1996] Stefan Langer. Selektionsklassen und Hyponymie im Lexikon. CIS-Bericht 96-94, Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximillians-Universität Mnchen, 1996.

[Laporte, 2005] E. Laporte. Graphes paramétrés et lexique-grammaire. In *Interface lexique-grammaire et lexiques syntaxiques et sémantiques*, 12 mars 2005.

[Maier-Meyer, 1995] Petra Maier-Meyer. Lexikon und automatische Lemmatisierung. CIS-Bericht 95-84, Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximillians-Universität München, 1995.

[Mallchok, 2004] Friederike Mallchok. *Automatic Recognition of Organization Names in English Business News*. Dissertation, Ludwig-Maximillians-Universität München, 2004.

[Mani *et al.*, 2000] Inderjeet Mani, Kristian Concepcion und Linda Van Guilder. Using summarization for automatic briefing generation. In *NAACL-ANLP 2000 Workshop: Automatic Summarization*. The MITRE Corporation, 2000.

[Mani, 2001] Inderjeet Mani. Recent developments in text summarization. In *Proceedings of the Tenth International Conference on Information und Knowledge Management*, pages 529–531, Atlanta, Georgia, USA, November 5-10 2001.

[Mann, 2001] Gideon S. Mann. Fine-grained proper noun ontologies for question answering. Technical report, Department of Computer Science, John Hopkins University, Baltimore, Maryland, 2001.

[MapPlanet GmbH, 2006] MapPlanet GmbH. MapPlanet.com, 2006. `http://mapplanet.com/ix/`.

[McCallum und Jensen, 2003] Andrew McCallum und David Jensen. A note on the unification of information extraction und data mining using conditional-probability, relational models. In *IJCAI'03 Workshop on Learning Statistical Models from Relational Data*, 2003.

[McCallum und Li, 2003] Andrew McCallum und Wei Li. Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction und Web-Enhanced Lexicons. In *Seventh Conference on Natural Language Learning (CoNLL)*, 2003.

[Mikolajewski, 2003] Tomasz Mikolajewski. Eine Untersuchung der Formen und Konstruktionen von Menschenbezeichnern für das Elektronische Lexikonsystem CISLEX. Studien zur Informations- und Sprachverarbeitung Band 7, Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximillians-Universität Mnchen, 2003.

[Miller *et al.*, 1993] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, und Katherine Miller. *Introduction to WordNet: An On-line Lexical Database*. Cognitive Science Laboratory, Princeton University, 1993.

[Mills, 2005] Elinor Mills. Google balances privacy, reach. *CNET News.com*, 14. Juli 2005. `http://news.com.com/Google+balances+privacy%2C+reach/2100-1032_3-5787483.html`.

[MoM, 2003] MoM. Ministry of Manpower - List of Occupations. Online-Verzeichnis, 2003. http://www.mom.gov.sg/MOM/MRSD/Others/2003W_OccList.pdf.

[Navarro *et al.*, 2001] Gonzalo Navarro, Ricardo Baeza-Yates und João Marcelo Arcoverde. Matchsimile: A flexible approximate matching tool for personal names searching. In *Proceedings SBBD '01*, pages 228–242, 2001.

[Nenkova und McKeown, 2003a] Ani Nenkova und Kathleen McKeown. References to named entities: A corpus study. In *Proceedings of NAACL-HLT*, 2003.

[Nenkova und McKeown, 2003b] Ani Nenkova und Kathleen McKeown. Improving the coherence of multi-document summaries: A corpus study for modeling the syntactic realization of entities. Technical report, Columbia University, 2003.

[Niu *et al.*, 2003] Cheng Niu, Wei Li, Jihong Ding und Rohini K. Srihari. Bootstrapping for named entity tagging using concept-based seeds. In *HLT-NAACL*, 2003.

[OOH, 1998] OOH. Occupational outlook handbook. Online-Verzeichnis, 1998. http://www.umsl.edu/services/govdocs/ooh9899/1.htm.

[OOH, 2000/2001] OOH. Occupational outlook handbook. Online-Verzeichnis, 2000-2001. http://www.umsl.edu/services/govdocs/ooh20002001/1.htm.

[Patel und Smarr, 2001] Stephen Patel und Joseph Smarr. Automatic Classification of Previously Unseen Proper Noun Phrases into Semantic Categories Using an N-Gram Letter Model. In *CS224N/Ling237 Final Projects*. Stanford University, 2001.

[Paumier, 2001] Sébastien Paumier. Some remarks on the application of a lexicon-grammar. *Lingvisticæ Investigationes XXIV:2*, pages 245–256, 2001. Amsterdam/Philadelphia, John Benjamins.

[Paumier, 2003] Sébastien Paumier. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Dissertation, Université de Marne-la-Vallée, 2003.

[Paumier, 2004] Sébastien Paumier. *Manuel d'utilisation d'Unitex*, 2004. http://wwwigm.univmlv.fr/~unitex/.

[Prospects.ac.uk, 2005] Prospects.ac.uk, 2005. http://www.prospects.ac.uk.

[Riloff und Jones, 1999] Ellen Riloff und Rosie Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteen National Conference on Artificial Intelligence (AAAI-99)*, pages 1044–1049, 1999.

[Roche, 1993] Emmanuel Roche. *Analyse syntaxique transformationnelle du français par transducteurs et lexique-grammaire*. Dissertation, Université Paris 7 - Denis Diderot, janvier 1993.

[Roth, 2002] Jeanette Roth. Der Stand der Kunst in der Eigennamen-Erkennung: Mit einem Fokus auf Produktenamen-Erkennung. Magisterarbeit, Universität Zürich, 2002.

[Schiffman *et al.*, 2001] Barry Schiffman, Inderjeet Mani und Kristian J. Concepcion. Producing biographical summaries: Combining linguistic knowledge with corpus statistics. In *Proceedings of the* 39^{th} *Annual Meeting of the Association for Computational Linguistics*, pages 450–457, Toulouse, France, 2001.

[Senellart, 1998a] Jean Senellart. Locating noun phrases with finite state transducers. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 1212–1219, Montréal, 1998.

[Senellart, 1998b] Jean Senellart. Tools for locating noun phrases with finite state transducers. In *The computational treatment of nominals. Proceedings of the Workshop, COLING-ACL'98*, pages 80–84, 1998.

[Senellart, 1999] Jean Senellart. *Outils de reconnaissance d'expressions linguistiques complexes dans des grands corpus*. Dissertation, Université Paris 7 - Denis Diderot, 1999.

[Silberztein, 1993] Max Silberztein. Dictionnaire électroniques et analyse automatique de textes - le système INTEX, 1993. Paris, Masson.

[Sparck-Jones, 1993] K. Sparck-Jones. What might be in a summary? In G. Knorz, J. Krause und C. Womser-Hacker, Hrsg., *Information Retrieval '93: Von der Modellierung zur Anwendung*, pages 9–26. Universitätssverlag Konstanz, 1993.

[SpecialistInfo.com, 2006] SpecialistInfo.com. Consultants, 2006. `http://www.specialistinfo.com/directory.php`.

[Sweeney, 2004] Latanya Sweeney. Finding Lists of People on the Web. *ACM Computers und Society*, 34(1), April 2004.

[Traboulsi, 2004] Hayssam Traboulsi. A local grammar for proper names. Magisterarbeit, University of Surrey, August 2004.

[Traboulsi, 2005] Hayssam N. Traboulsi. Towards the automatic acquisition of local grammars. In *Proceedings of the* 3^{rd} *Annual PhD Conference*, pages 13–18. Department of Computing, University of Surrey, UK, 2005.

[Tsur *et al.*, 2004] O. Tsur, M. de Rijke und K. Sima'an. Biographer: Biography questions as a restricted domain question answering task. In *Proceedings ACL 2004 Workshop on Question Answering in Restricted Domains*, 2004.

[U. S. Department of Labor, 2005] U. S. Department of Labor. Office of Administrative Law Judges. Internet Law Library, März 2005. `http://www.oalj.dol.gov/`.

[UBC, 1991] UBC. The University of British Columbia Library -1991 Standard Occupational Classification (SOC). Online-Verzeichnis, 1991. `http://data.library.ubc.ca/datalib/restricted/other/statscan/soc/alph1.html`.

[USDoL, 2001/2002] USDoL. U.S. Department of Labor - Alphabetical List of SOC Occupations. Online-Verzeichnis, 2001/2002. http://www.bls.gov/oes/2001/oes_alph.htm, http://www.bls.gov/oes/2002/oes_alph.htm.

[Waite, 2004] Maurice Waite, Hrsg. *Oxford Thesaurus of English*. Oxford University Press, 2004.

[Wikipedia, 2005/2006] Wikipedia. The Free Encyclopedia, 2005/2006. http://en.wikipedia.org/wiki/Main_Page.

[WordNet, Version 2.1] WordNet. An Electronic Lexical Database, Version 2.1. http://wordnet.princeton.edu/.

[Zhou *et al.*, 2004] Liang Zhou, Miruna Ticrea und Eduard Hovy. Multi-document biography summarization. In *Proceedings of EMNLP*, pages 434–441, 2004.

[Zoom Information Inc., 2006] Zoom Information Inc. ZoomInfo - People, Companies, Relationships, 2006. http://www.zoominfo.com/Search/.

Index

Abbreviation, 143
Abdankung, 132
Abkürzung, 70, 143
Abschluss, 104, 105
Adjektiv, 64, 118, 144
Adverb, 118
Akademischer Grad, 105
  Bachelor, 105
  Doktor, 105
  Master, 105
Akronym, 39, 81
Aktivform, 105, 109, 124, 125, 130, 132
Allgemeine Menschenbezeichner, 74
Altersangabe, 104
Anaphern, 77
Anredeform, 51–53
Antonym, 108
Antonymie, 53
Arbeitgeber, 125, 127
Arbeitnehmer, 127
Arbeitsbeziehung, 127
Arbeitsende, 129
Arbeitsverhältnis, 116, 125, 139
Aufwachsen, 100, 102
Australian Province Citizen, 56, 143
Auswertung, 135

Beruf, 124
Berufsausübung, 124
Berufsbezeichner, 24, 53–55, 73, 116, 118, 119, 126, 131, 134, 139, 143
Berufsbezeichnerlexikon, 38, 54
Berufsbezeichnung, 86
Berufung, 116
Beschäftigung, 125
Beschäftigungsverhältnis, 124, 126, 130
Beschäftigungszeit, 124
Bezahlung, 125
Biographie, 11
Biographische Relation, 11, 41, 45, 46, 56, 85, 97, 105, 137, 141
  Berufliche Relation, 115, 122
  Persönliche Relation, 11, 56, 97, 98, 111, 115
  Zufällige Relation, 12, 56
  Öffentliche Relation, 11, 12, 56, 115
Biography.com, 49
Biography.com Korpus, 46
Bootstrapping, 23, 24, 39, 67, 102
Borough, 61, 143
Bourough Adjective, 64, 143
Branche, 57, 121, 122, 144
Branchenlexika, 57
Breakeven-Punkt, 135
Bundesland, 62

Canadian Province, 61, 143
Canadian Province Citizen, 56, 143
Chomsky, Noam, 16
CISLEX, 48
CISLEX-E, 47
Citizen, 56, 143
City, 61, 143
City Adjective, 64, 143
Clustering, 34, 37
Company Name, 143
Continent, 61, 143
County, 61, 143

Département, 61, 62, 143
DAG, 16, 22
Datum, 91
Datumsangabe, 64, 69, 92, 94
Datumsangaben, 89
Datumserkennung, 64

DELA, 20
DELA Wörterbücher, 19
DELAC, 20, 35
DELACF, 20
DELAF, 20
DELAF-L, 21
DELAF-M, 21
DELAF-S, 21
DELAS, 20, 35
Directed Acyclic Graph, 16
Disambiguierung, 14, 15, 18, 72, 86
Diskursanalyse, 17
Drogenmissbrauch, 111

Ehejahre, 108
Ehepartner, 105, 107
Eigenname, 10, 24, 43, 48–50, 69, 73, 74, 77, 144
Eigennamenerkennung, 9, 10, 37
Einfache Wörter, 20
Einstellung, 122, 124
Einstellungsdatum, 124
Einwohner, 56, 143
Einwohnerlexikon, 55
Elektronisches Lexikon, 19
Eltern, 98, 100
Entität, 9, 10, 14, 39, 41, 43, 58, 69, 79, 97, 136, 137
Entlassung, 129
Ernennung, 116, 119
Ernennungsrelation, 116
Erziehung, 102
Evaluation, 136
Evaluationsmaß, 135
Extended Noun, 55, 61

Fachbereich, 143
Fachbereichlexikon, 57
Fachrichtung, 105
Fairon, Cédrick, 35
Fall-Out, 135
Financial Times, 38, 45, 46, 51
Firmeneintritt, 124, 125
Firmenname, 38, 58, 86, 143
Firmensitz, 124
First Name, 48, 49, 143
Flexionsform, 16, 19, 20
Friburger, Nathalie, 73
FT-Korpus, 46, 102, 117, 136
Full Name, 49, 50

Geburt, 98
Geburtsdatum, 98
Geburtsname, 100
Geburtsort, 98, 100, 102
Geburtstag, 98
Genauigkeit, 39, 47, 135
Geographical Term, 61, 143
Geographical Term Adjective, 64, 143
Geographischer Begriff, 85, 86
Gerichteter Azyklischer Graph, 16
Google, 67
Grafschaft, 61, 62, 143
grammatikalische Lexikoninformation
  `A`, 64
  `N`, 48–50, 54–56, 61
  `XA`, 64
  `XN`, 55, 61
Graph, 15, 16
Gross, Maurice, 18, 20, 25, 39, 56, 89

Harris, Zellig Sabbetai, 16
Heirat, 98, 105, 108
Herzversagen, 113
Himmelsrichtung, 86
Holonymie, 53
Homograph, 29
HPL, 38
Human, 48–50, 54–56, 143
Hyperonym, 41
Hyperonymie, 47, 53
Hyponym, 25
Hyponymie, 53

IE, 9
Information Retrieval, 42
Information Retrieval Methoden, 29
  Exact-Pattern-Algorithm, 29
  Exact-String-Matching-Algorithm, 29
  Key-Word-Algorithm, 29
  Statistical Algorithm, 29
Information-Retrieval-System, 135

Informationsextraktion, 9, 42
INTEX, 16, 18, 34

Jahresangabe, 95
Jahresspanne, 95
Jahresversammlung, 119
Jahreszahl, 92
Jahreszeiten, 94

Kanadische Provinz, 61, 143
Kanonische Form, 20
Kaskadierung, 34–36
Kategorietyp, 48–50, 54–56, 61, 64, 144
    grammatikalische Kategorie, 48
    semantische Kategorie, 48
Kindheit, 98, 100, 102
Klassifikation, 37
Komplexe Wörter, 20
Konkordanz, 22
Kontextfreie Grammatik, 16
Kontinent, 61, 62, 143
Kontraktionsauflösung, 19
Korpusverarbeitung, 18
Krankenhaus, 114
Krankheit, 111

LADL, 18, 20, 21
Land, 61, 62, 114, 144
Lehrbereich, 143
Lehreinrichtung, 104, 105
Lemma, 19
Lemmaform, 20
Lemmatisierung, 25, 56
Lexikalische Analyse, 19
Lexikon, 18
Lexikoneintrag, 47
Lexikongrammatik, 15, 18, 21
Lexikonkodierung, 21
Lexikonpriorität, 22
Lokale Grammatik, 14–18, 21, 22, 37, 38, 41–43, 46, 67, 69, 89, 97, 100, 104, 105, 111, 115, 131, 136, 141
Lokativa, 85, 86
Long Name, 49, 50, 144

Mallchok, Friederike, 41, 45, 53, 59
Mehrwortlexem, 21, 55, 61, 144
Menschenbezeichner, 14, 21, 41, 45, 48–50, 54–56, 69, 77, 97, 105, 122, 124, 126, 130, 132, 134, 137–139, 141, 143
Menschenbezeichnerlexika, 53
Meronymie, 53
Metropole, 61, 64, 143
Modularität, 115
Monatsabkürzung, 64, 92
Monatsangabe, 93
Monatsname, 64, 92
Morphologie, 16, 18
Mots Composés, 20
Mots Simples, 20
MUC-7, 10
Mustererkennung, 22

Nachfolge, 131
Nachfolgerelation, 117, 118
Nachname, 48, 49, 73, 144
    einfacher Nachname, 48
    komplexer Nachname, 48
Nachnamenlexikon, 48
Nagel, Sebastian, 45
Named Entity, 9–11
Named Entity Recognition, 9, 37, 41, 42
Namen der Lexika
    Citizens-.dic, 55
    Companies-.dic, 59
    Disciplines-.dic, 57
    FirstNames-.dic, 48, 49
    GeosMapplanet-.dic, 61, 62
    GeosSebastian+.dic, 61
    GeosWikipedia-.dic, 61, 62
    HumVP-.dic, 56
    JD-.dic, 54
    LastNames-.dic, 48
    LongNamesAuthors-.dic, 51
    LongNamesBios-.dic, 49
    LongNamesFT-.dic, 51
    LongNamesSpecialistInfo-.dic, 51
    MenbezWordnet-.dic, 54
    Month-.dic, 64
    MonthAbbr-.dic, 64

Titles-.dic, 52
USstates-.dic, 63
context_before-.dic, 60
org-.dic, 59
org_adj-.dic, 60
orgbez-.dic, 59
Nation, 61, 144
Nation Adjective, 64, 143
NER, 9, 10, 42, 43
Neubesetzung, 131
New York City Bourough, 61, 144
New York City Citizen, 56, 144
New York Times, 38
Nominalphrase, 97, 118, 119
Normalisierung, 19, 46

ODL, 38, 59
ONL, 38, 58, 59
Organisation, 41, 79, 81, 83, 122, 124, 125, 130, 131, 137
Organisationsbeschreibungslexikon, 38, 58
Organisationsname, 37, 38, 58, 69, 72, 79, 81, 86, 131, 137, 143
Organisationsnamenlexikon, 38, 58, 59
Ortsangabe, 85, 86, 114
Ortsbezeichnung, 38

Pars Pro Toto, 70
Passivform, 105, 109, 124, 125, 130, 132
Paumier, Sébastien, 18, 19
Pensionierung, 134
Performanz, 39, 47
Person, 124, 137
Personenbezogene Prädikate, 56, 137, 138, 141
Personenlexikon, 52
Personenname, 49–52, 69, 72, 79, 81, 83, 124, 144
Personennamenlexika, 49, 73
Personensuchmaschine, 51
Prädikat, 98, 116
Precision, 36, 37, 135, 136
Prenom-prolex, 35
Prioritätsebene, 22
Produktname, 43
Prolintex, 35
Proper Noun, 48–50, 144
Province Adjective, 64, 143

Recall, 36, 37, 135, 136
Region, 61, 62, 144
Rekursiver Aufruf, 91
Relative Jahresangabe, 93
Relativsatz, 97
Rentenalter, 134
Ressourcen, 45
Reuters Korpus, 38, 42, 45
Reuters News, 38
Roth, Jeannette, 43

Satzenderkennung, 19, 34, 46
Satzendmarkierung, 46
Scheidung, 98, 108, 110
Scheinname, 72
Schulabschluss, 98, 102
Schule, 104
Sektor, 57, 121, 144
Sektorenlexikon, 57
Semantik, 16, 21
semantische Lexikoninformation
`ABourough`, 64, 143
`ACity`, 64, 143
`AGEO`, 64, 143
`ANation`, 64, 143
`AProvince`, 64, 143
`AState`, 64, 143
`Abbrev`, 143
`AuProvinceCitizen`, 56, 143
`Borough`, 61, 143
`CaProvinceCitizen`, 56, 143
`CaProvince`, 61, 143
`Citizen`, 56, 143
`City`, 61, 143
`Company`, 143
`Continent`, 61, 143
`County`, 61, 143
`DayOfWeek`, 65
`Departement`, 61, 143
`Discipline`, 143
`FN`, 48, 143
`GEO`, 61, 143
`HumVP`, 57

Hum, 48–50, 54–56, 143
JD, 55, 143
LN, 50, 144
NYCBourough, 61, 144
NYCcitizen, 56, 144
Nation, 61, 144
Ntime, 65
PR, 48–50, 144
Region, 61, 144
SN, 49, 144
Sector, 144
Title, 144
USstateCitizen, 56, 144
USstate, 61, 144
Urbanite, 56, 144
Senellart, Jean, 34, 39
SpecialistInfo.com, 51
Sprachgebundenheit, 42
Städtenamen, 62
Staatsbürger, 56, 143
Stadt, 61, 64, 114, 143
Stadtbewohner, 56, 144
Stadtteil, 61, 143
State Adjective, 64, 143
Sterben, 111
Straßenname, 102
Subgraph, 16
Subsprache, 17, 38
Suchfenster, 137
Surname, 49, 144
Sweeney, Latanya, 51
Synekdoche, 70
Synonym, 25, 98, 115, 119
Synonymie, 53, 110
Synset, 53
Syntaktische Variabilität, 13, 69, 79
Syntax, 16, 18, 21

Tätigkeit, 126
Tagesangabe, 92, 94
Teilformenlexikon, 21
Temporalialexikon, 64
Textkodierung, 18
Titel, 51, 52, 144
Adelstitel, 52
akademischer Grad, 51, 52
aristokratischer Titel, 52
militärischer Rang, 51, 52
Titellexikon, 52
to be appointed, 116, 122
to be born, 98, 100
to be dismissed, 129
to be divorced, 108
to be employed, 124
to be married, 105
to be paid, 125
to be raised (up), 98, 100
to be replaced, 131
to die, 111
to dismiss so., 129
to employ so., 122
to graduate, 102
to join, 124
to resign, 132
to work, 126
Tod, 98, 111
Todesalter, 111
Todesdauer, 111
Todesort, 114
Todesursache, 111, 113
Todeszeitpunkt, 111
Tokenisierung, 19, 21, 46
Tokenizer-Programm, 45, 46
Toponym, 10, 35, 41, 43, 61–63, 69, 72, 85, 86, 143
Toponymlexikon, 61
Transduktor, 15, 22, 34–36, 39
Transitionsnetz, 93, 115
Trennung, 110

Unfall, 113
unique beginners, 54
UNITEX, 16, 18, 22, 34, 45, 46
US Bundesstaat, 61, 63, 144
US State, 61, 63, 144
US State Citizen, 56, 144

Verbalphrase, 98, 100, 116, 124
Verbalphrasenlexikon, 56
Verben
einfache Verben, 56

komplexe Verben, 56
Verifikation, 67
Vollform, 70
Vollständigkeit, 135
Vorname, 48, 49, 69, 73, 143
einfacher Vorname, 48
komplexer Vorname, 48
Vornamenlexikon, 48

Wörterbuch, 15, 47, 50, 51
Wörterbuch-Lookup, 85
Wörterbucheintrag, 20, 21, 41, 49, 50, 67, 79, 83, 122, 136
Wall Street Jounal, 38
Wikipedia, 52, 54, 55, 57
Wirtschaftsnachrichten, 14, 42, 45, 46
Wochentag, 65, 91
Wochentaglexikon, 65
WordNet, 54

Zeitangabe, 38, 64, 91
Zeitbestimmung, 64
Zeitbezogene Nomina, 65
Zeitspanne, 93