Ingeborg Blank
Centrum für Informations- und Sprachverarbeitung
der Ludwig-Maximilians-Universität München
Oettingenstr. 67
80538 München

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Philosophie
an der Ludwig Maximilians-Universität
zu München

1997

ISBN-Nr. 3-930859-10-6

Meinen Eltern

Vorwort

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit am Centrum für Informations- und Sprachverarbeitung der Universität München.

Während meines Studiums der Linguistik und Informatik und vor allem während eines längeren Studienaufenthalts in Frankreich wurde ich immer wieder mit Fragen der Übersetzung konfrontiert. Meine ersten Kontakte zu professionellen Übersetzern konnte ich bei zwei Praktika in einem technischen Übersetzungsdienst knüpfen. Dort bestand meine Aufgabe im wesentlichen darin, die Beschreibungen technischer Änderungen in Dokumenten zu übersetzen, wobei ich mich auf bestehende Übersetzungen stützte.

Das Interesse für Fremdsprachen und Übersetzungen blieb mir auch nach meiner Rückkehr nach Deutschland erhalten, als ich nach Abschluß des Studiums meine erste Arbeitsstelle in einem großen Wörterbuchverlag antrat.

Herzlich danken möchte ich Herrn Prof. Dr. Guenthner, der die vorliegende Arbeit betreut hat und der mich immer wieder durch neue Anregungen und hilfreiche Ratschläge unterstützt hat.

Danken möchte ich auch den Mitarbeitern des Europäischen Patentamts in München und Wien, die mir freundlicherweise für meine Untersuchungen ein großes mehrsprachiges Fachtextkorpus zur Verfügung gestellt haben.

Elke Sagenschneider und Gabi Schicht verdanke ich konstruktive Vorschläge und Diskussionen bei der Erstellung und Korrektur des Manuskripts.

Und last but not least möchte ich mich bei meinem Mann und bei meiner Familie für ihre Unterstützung bedanken.

München, den 18. März 1997

Ingeborg Blank

Inhaltsverzeichnis

1 Einleitung

Im Zusammenhang mit Schlagwörtern wie “Technologietransfer”, “internationale Zusammenarbeit” und “Explosion des Wissens” wird immer wieder auf die zunehmende Bedeutung fachsprachlicher Übersetzungen hingewiesen. Allein in Europa rechnet man am Ende dieses Jahrtausends mit einem Übersetzungsvolumen von etwa 1,8 Mio. Seiten pro Jahr[1], was einem Markt von mehr als 70 Milliarden US Dollar entspricht. Nicht zuletzt bedingt durch die europäische Norm, die die Übersetzung von Bedienungsanleitungen in die jeweilige Sprache des Absatzmarkts vorschreibt, haben Dokumente aus dem Bereich der technischen Dokumentation wie Bedienungsanleitungen, Reparaturanleitungen und Handbücher den größten Anteil am Übersetzungsvolumen.

Aufgrund des wachsenden Bedarfs an Fachübersetzungen ist es verständlich, daß gegenwärtig ein starkes Interesse an Maschineller Übersetzung bzw. an effizienten Werkzeugen zur Unterstützung des Übersetzungsprozesses besteht. Vollautomatisch erstellte, qualitativ hochwertige Übersetzungen können jedoch mit den bisherigen Systemen allenfalls für Texte sehr begrenzter Fachgebiete angefertigt werden.

Eine Sammlung von Texten und deren Übersetzungen stellt eine wichtige Informationsquelle für einen Übersetzer dar. Ein solches Textkorpus bedarf aber einer entsprechenden Aufbereitung, um seinen vollen Nutzen entfalten zu können. P. Isabelle hat diesen Gedanken in einem Artikel, der sich mit den ergonomischen Aspekten der Übersetzung beschäftigt, folgendermaßen formuliert:

> Given the staggering volume of translations produced year after year, it is quite obvious that existing translations contain more solutions to more translation problems than any other existing resource. Unfortunately translators can currently derive very little benefit of this fact[2].

Diese Idee war der Ausgangspunkt für die vorliegende Arbeit. Auf der Grundlage eines Korpus von Fachtexten und ihren Übersetzungen erschien eine terminologische Untersuchung von besonderem Interesse. Das Ziel dieser Untersuchung besteht letztlich in der Entwicklung eines Verfahrens, mit dem ein Fachtextkorpus für die Erstellung von mehrsprachigen Glossaren zumindest halbautomatisch aufbereitet werden kann. Glossare sind eine Grundlage für die Erstellung von Wörterbüchern und werden besonders dann gebraucht, wenn aus Konsistenzgründen eine Übersetzung standardisiert werden soll.

Ein dreisprachiges (dt.-engl.-frz.) Textkorpus einer Größe von 12 Mio. Wörtern aus dem Bereich der Patentdokumentation stand zur Verfügung. Die Aufbereitung sollte es ermöglichen, Vorkommen möglicher Fachtermini aus den Daten zu extrahieren und

1. Schmitt, 1993; die zitierten Schätzungen sind das Ergebnis einer Umfrage unter Übersetzern und Dolmetschern.

2. Isabelle 1992, S. 8.

nach Auffinden eines relevanten Beispiels in einem Quelltext die entsprechende Stelle im Zieltext zu lokalisieren. Dies erfordert erstens eine Alignierung der Texte mit ihren Übersetzungen auf Satzebene und zweitens für jede Einzelsprache eine Definition von Fachtermini mit formalisierbaren Kriterien, d. h. eine Definition, die in einem automatischen Verfahren umsetzbar ist. Eine solche Definition liegt beispielsweise dann vor, wenn ein Fachterminus als Nominalphrase mit sprachspezifischen Bildungsmustern definiert wird wie z. B. im Deutschen als Nominalkompositum (*Beschwerdeverfahren*), im Englischen als eine Nomen-Nomen-Verbindung (*appeals procedure*) und im Französischen als Verbindung aus einem Nomen und einer Präpositionalgruppe (*procédure de recours*).

Zwei Implementierungen von Alignierungsverfahren wurden für alle Sprachpaare des Korpus evaluiert. Da die Extraktion von Termini eng an die einzelsprachlichen Besonderheiten der Wortbildung gebunden ist, wurde im Rahmen einer Untersuchung, die den Schwerpunkt mehr auf den qualitativen als auf den quantitativen Aspekt legt, die Extraktion von Termini auf das Sprachpaar Deutsch-Französisch begrenzt.

In den letzten Jahren ist das Interesse an Untersuchungen von Fachtexten mit dem Ziel einer automatischen Terminologieextraktion gewachsen. Einerseits liegt ein immer größerer Anteil der Fachdokumentation in elektronischer Form vor; andererseits können die in der maschinellen Sprachverarbeitung entwickelten Techniken immer größere Textmengen analysieren.

In der Linguistik und Computerlinguistik ist ein neuer Trend zu beobachten. Empirische Methoden, die in den 50er Jahren populär waren, wurden “wiederentdeckt” und anhand großer Textkorpora evaluiert und weiterentwickelt. Dieser Ansatz hat eine Reihe von Produkten und Technologien hervorgebracht. Dazu gehören Verfahren, mit denen die Wortformen eines Textes automatisch mit Wortklassen annotiert werden (*POS-Tagging*)[1] oder Alignierungsverfahren für mehrsprachige Texte. Diese Verfahren können jetzt für die Lösung anderer linguistischer Probleme eingesetzt werden.

Arbeiten zur mehrsprachigen Terminologieextraktion sind, wie gezeigt, die Grundlage für die Entwicklung effizienter Hilfsmittel für die Fachtextübersetzung, zu denen in jüngster Zeit neben den bereits erwähnten Glossaren und Terminologiedatenbanken auch Übersetzungsspeicher gehören. Übersetzungsspeicher (*translation memories*) sind Programme, die in Ausgangstexten bereits übersetzter Dokumente nach identischen oder ähnlichen Textsegmenten suchen und die entsprechenden zielsprachlichen Übersetzungen bei der Bearbeitung eines neuen Dokuments als Übersetzungsvorschlag anbieten. Die Systeme bedienen sich zur Ermittlung identischer oder ähnlicher Dokumente unterschiedlicher Parsingalgorithmen sowie der *Fuzzy Logic* (Unscharfe Logik), wobei der Benutzer z. T. den Grad der Übereinstimmung von Sätzen, die bei der Übersetzung berücksichtigt werden sollen, selbst angeben kann.

1. POS ist eine Abkürzung für *part of speech* (Wortklasse).

Mehrsprachige Terminologie wird auch bei der Entwicklung und Anpassung von Systemen zur Maschinellen Übersetzung für neue Fachgebiete benötigt.

Arbeiten zur Terminologieextraktion sind nicht nur unter bilingualen sondern auch unter monolingualen Aspekten von Bedeutung. Die Dokumentationswissenschaft verzeichnet einen wachsenden Bedarf an terminologischen Resourcen. Die thematische Klassifizierung von Dokumenten (Indexierung) erfolgt über die Zuordnung von Deskriptoren, die in den meisten Fällen Termini eines Gebiets sind.

Komposita machen in vielen Sprachen den größten Teil der Terminologie aus. Unter diesem Aspekt können Arbeiten zur Extraktion von Terminologie neue Erkenntnisse zur automatischen Erkennung von Komposita liefern. In der Praxis braucht jedes *NLP*-System (*natural language processing system*: System, das natürliche Sprache verarbeitet) Lexikonkomponenten für die Kompositaerkennung und -analyse.

Überblick

Der folgende Haupteil der Arbeit gliedert sich in sechs Kapitel. Zuerst soll geklärt werden, mit welchen Kriterien Fachtexte in der Fachsprachenforschung und in der Übersetzungswissenschaft definiert werden. Darauf folgt eine Beschreibung übersetzungsorientierter Terminologiearbeit in Theorie und Praxis. Es werden einige computerlinguistische Arbeiten und Projekte vorgestellt, die sich mit der Entwicklung von Hilfsmitteln für die praktische Terminologiearbeit beschäftigen. Ein Überblick über den Stand der Technik in der automatischen Terminologieextraktion war die Grundlage für die Entwicklung eines eigenen Verfahrens.

Die automatische Lokalisierung von Fachtermini in Texten erfordert - außer einer Alignierung der Texte auf Satzebene - eine Vorverarbeitung der Texte (Lemmatisierung und POS-Tagging) und die eigentliche Extraktion der Vorkommen der definierten Suchmuster. Danach muß geprüft werden, ob die für jede Einzelsprache extrahierten Sequenzen linguistisch korrekte Nominalphrasen sind, um in einem zweiten Schritt der Frage nachzugehen, ob es sich inhaltlich und funktional um Fachterminologie handelt. Diese zweite Frage kann meines Erachtens letztlich nur von einem Terminologen beantwortet werden. Eine Auswertung der extrahierten Sequenzen nach statistischen Kriterien kann allenfalls Hinweise auf mögliche Fachtermini geben.

Im dritten Kapitel wird das Textkorpus vorgestellt, das für die vorliegende Arbeit zur Verfügung stand. Es handelt sich um ein Textkorpus aus der Patentdokumentation. Zum besseren Verständnis der angeführten Textbeispiele werden in einem Exkurs die inhaltlichen Grundlagen der Patentdokumentation erläutert. Daran schließt sich eine Beschreibung der Strukturierung der einzelnen Dokumente an, auf deren Grundlage Teilkorpora erstellt werden konnten.

Im vierten Kapitel werden verschiedene Satzalignierungsverfahren diskutiert. Für die beiden gängigsten Algorithmen (die satzlängenbasierte Methode nach Church und Gale und die linguistisch basierte Methode nach Kay und Röscheisen) lagen Imple-

mentierungen vor. Es wurden mehrere Testreihen mit Teilkorpora unterschiedlicher Sprachrichtungen durchgeführt. Die abschließende Evaluierung zeigt, für welche Textsorten welcher Algorithmus besser geeignet ist und gibt Anregungen für mögliche Verbesserungen und Weiterentwicklungen.

Das fünfte Kapitel ist der Definition und Extraktion potentieller Fachtermini im Französischen gewidmet. Ein Überblick über die linguistische und computerlinguistische Forschung zur Kompositabildung im Französischen ist der Ausgangspunkt für eine Definition nominaler Fachterminologie. Eine Definition potentieller Fachtermini als Nominalphrasen bestimmter Bildungsmuster ist in den Fällen problematisch, in denen unklar ist, ob eine konkrete komplexe Nominalphrase als eine Einheit zu betrachten ist oder weiterzerlegt werden muß. Diese Problematik wird an Hand von Beispielen aus dem Korpus und aus der Literatur erörtert. Die eigentliche Extraktion wurde mit dem INTEX-System[1] und eigenen Programmen durchgeführt. Im Vorfeld wurde untersucht, wie sich das INTEX-System für diese Aufgabe eignet und in welchen Punkten eine Anpassung erforderlich und auch möglich ist. Zum Schluß werden die Ergebnisse der Extraktion dargestellt und evaluiert.

Das sechste Kapitel behandelt die Definition und Extraktion potentieller Fachtermini im Deutschen. Auch hier waren linguistische und computerlinguistische Arbeiten Ausgangspunkt für eine Begriffsbestimmung. Als Informationsquellen wurden einschlägige Literatur aus der Fachsprachenforschung sowie Ergebnisse eines Forschungsprojekts zur Patentdokumentation herangezogen. Nominalkomposita machen den größten Teil der Fachtermini aus, Nominalphrasen anderer Bildungsmuster spielen eine untergeordnete Rolle, wurden aber bei der Extraktion ebenfalls berücksichtigt. Für die Lemmatisierung und das POS-Tagging der Texte stand das CISLEX-System[2] zur Verfügung, dessen Arbeitsweise kurz skizziert wird. Abschließend werden die Ergebnisse der Extraktion dargestellt und ausgewertet.

Im letzten Kapitel wird eine kontrastive Auswertung vorgenommen und eine Anwendung dargestellt. Die Untersuchung möglicher Übersetzungsäquivalenzen berücksichtigt zwei Gesichtspunkte: die Übersetzungen einzelner Fachtermini und die Strukturäquivalenzen in den beiden Sprachrichtungen. Die Ergebnisse werden in Zusammenhang mit der in der Literatur oft vertretenen Hypothese der 1:1-Entsprechungen von Fachterminologie in verschiedenen Sprachen ausgewertet. Als mögliche Anwendung wird ein Konkordanzprogramm vorgestellt, das Informationen zu einzelnen Fachtermini und zu speziellen Verwendungskontexten bietet sowie eine Zusammenstellung von Termini mit gemeinsamen Konstituenten ermöglicht.

1. Silberztein 1993.
2. Maier-Meyer 1995.

2 Fachsprache und Übersetzung

Das Ziel der vorliegenden Arbeit ist eine computerlinguistische Untersuchung von Fachtexten und deren Übersetzungen unter terminologischen Aspekten. Ich halte es für wichtig darzustellen, wie Terminologie in der Fachsprachenforschung und in der Übersetzungswissenschaft definiert wird und wie sich Terminologiearbeit in der Theorie und in der Praxis gestaltet. Die Anforderungen, die an Terminologie gestellt werden, erscheinen auf den ersten Blick unvereinbar mit einem Ansatz, der nicht auf semantischen Kriterien basiert, was bei einem computerlinguistischen Ansatz meistens der Fall ist. Darstellungen aus der Praxis der Terminologiearbeit zeigen aber, daß die Ergebnisse "computerlinguistischer Terminologiearbeit" im realen Übersetzungskontext durchaus von Nutzen sein können.

Es werden einige computerlinguistische Arbeiten sowie ein konkretes Forschungsprojekt vorgestellt und diskutiert. Sie waren die Grundlage für eigene inhaltliche und methodische Überlegungen.

2.1 Fachsprachenforschung und Übersetzungswissenschaft

2.1.1 Gemeinsprache und Fachsprachen

"Fachsprache" ist eine häufig verwendete Bezeichnung, die alle Formen der fachbezogenen Kommunikation meint und oft in Gegensatz zum Begriff der üblicherweise verwendeten Gemein- oder Standardsprache gestellt wird.

Zum Begriff der Fachsprache und seiner Eingrenzung hat die Sprachforschung wichtige grundsätzliche Überlegungen angestellt. Eine einheitliche Fachsprachendefinition liegt aber bisher noch nicht vor[1]. Die verschiedenen Definitionen betonen die Verwendungszusammenhänge von Sprache und machen deutlich, daß Fachsprachen keine eigenen Sprachsysteme sind, sondern vielfältige Beziehungen zur Gemeinsprache besitzen. Außerdem zeigen die Definitionen in ihren allgemein gehaltenen Bestimmungen, daß es eine vollbefriedigende Definition von Fachsprache nicht geben kann und es *die* Fachsprache nicht gibt.

In Hoffmanns Definition steht der kommunikative Aspekt im Vordergrund:

> Fachsprache - das ist die Gesamtheit aller sprachlichen Mittel, die in einem fachlich begrenzbaren Kommunikationsbereich verwendet werden, um die Verständigung zwischen den in diesem Bereich tätigen Menschen zu gewährleisten.[2]

1. Definitionen u. a. durch Möhn & Pelka 1984, Drozd & Seibicke 1973, Fluck 1985, Hoffmann L. 1987.

2. Hoffmann L. 1985, S. 170.

Möhn betont in seiner Definition, daß Fachsprache immer ein Teil der Gemeinsprache ist:

> Fachsprachen bilden innerhalb der Gesamtsprache auf einzelne Fachgebiete bezogene, in sich differenzierte Subsysteme, die durch eine charakteristische Auswahl, Verwendung und Frequenz sprachlicher Mittel definiert sind.[1]

Die meisten Autoren unterscheiden eine **horizontale** und eine **vertikale Gliederung** der Fachsprachen. Die horizontale Gliederung bezieht sich auf die Einteilung des Wissens in die einzelnen Fachgebiete. Unter vertikaler Gliederung versteht man die (oft problematische) Unterteilung der Fachsprache eines Fachgebiets in verschiedene Sprachebenen.

2.1.2 Charakteristika der Fachsprachen

Die Frage, in welchen konkreten Punkten sich fachsprachliche Texte von gemeinsprachlichen unterscheiden, läßt sich nur schwer pauschal beantworten. Der Grund liegt in der bereits angedeuteten Vielschichtigkeit des Phänomens Fachsprache. Untersuchungen zu den Charakteristika der Fachsprachen sind u. a. von den Vertretern der funktionalen Stilistik[2] auf der Grundlage von Fachtexten der wichtigeren europäischen Sprachen durchgeführt worden. Dabei wurde eine Reihe von Merkmalen erarbeitet, die allen oder einer großen Anzahl von Fachsprachen gemeinsam sind; die wichtigsten von ihnen lassen sich in Anlehnung an Hoffmann[3] folgendermaßen zusammenfassen:

1. Das Verb verliert seinen konkreten Zeitbezug und steht zumeist im Präsens, insbesondere in der 3. Person Singular.
2. Das Verb steht häufig im Passiv.
3. Das Verb als Wortart spielt eine geringe Rolle.
4. Das Substantiv spielt eine wichtige Rolle.
5. Der Singular ist wesentlich häufiger als der Plural.
6. Das Adjektiv tritt verhältnismaßig häufig auf.

Hoffmann hat auch statistische Untersuchungen zu den lexikalischen Charakteristika von Fachsprachen[4] durchgeführt. Diese Untersuchungen bestätigen die bedeutende Rolle der Substantive, die in dt. Fachtexten bis zu 44 % des gesamten Wortschatzes gegenüber 28 % in gemeinsprachlichen Texten ausmachen können. Adjektive mit Differenzierungsfunktion erreichen in deutschen Fachtexten eine Frequenz bis zu 16,2 % (gegenüber 9,6 % in der Gemeinsprache). Demgegenüber spielt das Verb ebenso wie das Adverb in der Fachsprache eine wesentlich geringere Rolle als in der Gemeinspra-

1. Möhn & Pelka 1984, S. 175.
2. Z. B. Drozd & Seibicke 1973, S. 86 f.
3. Hoffmann L. 1985, S. 238 f.
4. Hoffmann L. 1985, S. 274 f.

che. Hoffmann faßt dies in der Regel zusammen, daß die Zahl der Nomina in einem Text umgekehrt proportional zur Zahl der Verben ist[1]. Von Interesse ist auch, daß ein von Hoffmann vorgenommener Vergleich der Verteilung der Wortarten in verschiedenen Fachgebieten eine weitgehende Übereinstimmung aufzeigte. Linguistische Untersuchungen dieser Art stützen somit die Auffassung, daß es grundlegende Unterschiede zwischen Fachsprache und Gemeinsprache gibt.

Das Fachwort (=Terminus) ist für die Fachsprache von zentraler Bedeutung, wenngleich es nicht das einzige Wesensmerkmal von Fachsprache ist.

2.1.3 Der Fachterminus: Problematik seiner Charakterisierung

2.1.3.1 Die Definition des Terminus in der Fachsprachenforschung

Charakterisierungen von Fachtermini und Angaben zu ihrer Identifizierung sind in der Literatur reichlich zu finden. Trotz verschiedener definitorischer Ansätze scheint über ihre Funktion und Kennzeichen weitgehend Einigkeit zu herrschen. Die Definitionen unterscheiden sich allein darin, wie umfassend sie sind bzw. welche Kriterien der Fachsprachlichkeit sie jeweils stärker gewichten. So findet man eine Reihe recht präzise gehaltener Beschreibungen und Definitionen, die zumeist mit dem Anspruch auf universelle Gültigkeit sowohl für alle Einzelsprachen als auch für alle Fachsprachen auftreten.

Die wichtigsten, von den Sprachwissenschaftlern übereinstimmend genannten Charakteristika werden von Fluck folgendermaßen zusammengefaßt:

> Gegenüber den gemeinsprachlichen Wörtern zeichnen sich die Fachwörter vor allem durch ihren fachbezogenen Inhalt und ihre Kontextautonomie aus. Als weitere Eigenschaften werden in der Literatur die Tendenz zu Exaktheit, Eindeutigkeit, Begrifflichkeit, Systematik, Neutralität und Ausdrucksökonomie genannt.[2]

Die hier von Fluck angeführten Kriterien waren ursprünglich von L. Hoffmann[3] zusammengetragen worden. Sie gehen zum Teil auf die Ausführungen von Eugen Wüster, dem Begründer der Allgemeinen Terminologielehre[4], zurück und sind von fast allen Fachsprachenforschern aufgegriffen und ergänzt worden.

2.1.3.2 Eigenschaften des Fachterminus

Unter **Kontextautonomie** ist die Unabhängigkeit des Fachterminus von seiner sprachlichen Einbettung zu verstehen (“der Terminus braucht keinen Kontext, um verstanden

1. Hoffmann L. 1985, S. 278.
2. Fluck 1985, S. 33.
3. Hoffmann L. 1985, S. 308.
4. Wüster 1970.

zu werden"[1]). **Exaktheit** steht für die genaue Bedeutungsfestlegung des Fachworts und seine Abgrenzung gegenüber anderen Fachwörtern. **Eindeutigkeit** (z. T. auch als Monosemantizität bezeichnet) meint, daß das Fachwort jeweils nur auf eine bestimmte fachliche Erscheinung bezogen ist, also nur einen fachlichen Begriff repräsentiert. **Begrifflichkeit** bezeichnet die Einbindung des Fachworts in ein Begriffssystem und seine Funktion als sprachliches Zeichen für eine gedankliche Einheit, den Begriff. In engem Zusammenhang damit steht das Merkmal der **Systematik**, das die Beziehung des Begriffs zu anderen Begriffen, d. h. seine Einbettung in ein Begriffssystem, zum Ausdruck bringt. Stilistische **Neutralität** weist auf die Rationalität und Objektbezogenheit der fachsprachlichen Verständigung hin, die sich im weitgehenden oder völligen Fehlen ästhetischer, expressiver und modaler Komponenten im Fachwortschatz äußert. **Ausdrucksökonomie** leitet sich aus dem Bestreben nach fachlicher Präzision und nach formaler Kürze und Knappheit ab.

Noch ein weiteres Kriterium wird häufig zur Abgrenzung des fachsprachlichen vom gemeinsprachlichen Wortschatz herangezogen[2]. Im Unterschied zu den Gemeinsprachen, die die außersprachliche Wirklichkeit auf unterschiedliche Weise gliederten, folgten die Fachsprachen einer einheitlichen, durch die Sachen selbst unzweideutig vorgegebenen Strukturierung. Die Konsequenz hieraus sei die vollständige zwischensprachliche Kongruenz und damit einhergehend die 1:1-Substituierbarkeit der Fachtermini[3].

Kritik

Wenn die Fachterminologien aller Einzelsprachen sämtliche soeben skizzierte Kennzeichen aufwiesen und die fachlichen Begriffe zudem über die Sprachgrenzen hinweg dieselben wären, dann dürften sich die zwischensprachlichen Unterschiede auf einige formale Aspekte (z. B. Wortbildung) beschränken. Tatsächlich werden die Fachterminologien unter Hinweis auf diesen Umstand von einigen Sprachwissenschaftlern aus der strukturellen Untersuchung des Wortschatzes auch explizit ausgeklammert. Demgegenüber steht eine Fülle von Publikationen aus dem Gebiet der Fachübersetzung, in denen Probleme der fachsprachlichen Lexik angesprochen werden.

2.1.4 Terminologiearbeit in der Übersetzungswissenschaft

Terminolegiearbeit kann unter zwei Aspekten betrieben werden: normativ oder deskriptiv.

Die **normative** Terminologiearbeit hat das Ziel, einen Soll-Zustand von Fachsprache zu definieren, um durch Vermeidung von Mehrdeutigkeiten und Ausdrucksvarianten

1. Hoffmann L. 1985, S. 163.

2. U. a. von Coseriu 1966.

3. Diese These wird u. a. in der Arbeit von Reinart 1993 für das Sprachpaar Deutsch-Französisch widerlegt.

fachsprachliche Kommunikation zu erleichtern. Die normative Terminologiearbeit postuliert Eindeutigkeit von Benennungen und beschreibt damit eine Idealvorstellung, die durch die Postulate der Allgemeinen Terminologielehre beeinflußt ist.

Wird die normative Terminologiearbeit[1] mehrsprachig betrieben, so geht man von einer interlingualen Äquivalenz von Benennungen aus, d. h. von der Annahme, daß einer bestimmten Benennung eines Begriffs in der einen Sprache auch eine bestimmte Benennung desselben Begriffs in der anderen Sprache entspricht. Die fachsprachliche Realität sieht jedoch anders aus.

In der **deskriptiven** Terminologiearbeit geht es um die Beschreibung des bestehenden Sprachzustands:

> Die *deskriptive* Terminologiearbeit erfaßt den *Ist*-Zustand der Fachsprache und ist mithin die adäquate Methode für jene, die mit den vielfältigen und in mancher Hinsicht oft defekten Erscheinungsformen fachsprachlicher Texte konfrontiert werden wie etwa Übersetzer und Dolmetscher. Eine am Informationsbedarf von Übersetzern/Dolmetschern (kurz: translationsorientierte) Terminologiearbeit muß *deskriptiv* sein[2].

Übersetzungsbezogene Terminologiearbeit[3] wird in der Regel von Terminologen und Übersetzern durchgeführt, die sowohl eine sprachliche bzw. fachsprachliche als auch eine fachliche Vorbildung haben. Werden ein Sachgebiet und seine Terminologie im Zusammenhang bearbeitet, so spricht man von systematischer Terminologiearbeit. Die berufliche Wirklichkeit läßt dies allerdings nicht immer zu. Daher spielen “punktuelle” Untersuchungen häufig eine weit wichtigere Rolle, d. h. der Terminologe bemüht sich, einzelne Wortgleichungen möglichst mit einem relevanten, d. h. aussagekräftigen Kontext zu finden, die der Fachübersetzer für eine konkrete Übersetzung benötigt.

Die systematische Bearbeitung mehrsprachiger Terminologien umfaßt im wesentlichen die folgenden Arbeitsschritte:

- Abgrenzung des Fachgebiets
- Aufteilung des Fachgebiets in kleinere Einheiten
- Beschaffung und Analyse des Dokumentationsmaterials für die zu bearbeitenden Sprachen
- Sammlung und vorläufige Zuordnung der gefundenen Benennungen und Begriffe sowie aller zweckdienlichen Informationen
- Erarbeitung der einzelsprachlichen Begriffssysteme
- terminologische Analyse

1. Vgl. dazu Felber & Budin 1989.
2. Schmitt 1994, S. 33.
3. Vgl. dazu Arntz & Picht 1982.

- Bestimmung und Erklärung von möglichen Übersetzungsäquivalenten bzw. Begriffsinkongruenzen[1].

Diese hohen Anforderungen an die Qualität der Terminologie können in der Praxis jedoch oft nicht erfüllt werden, wie aus Berichten über den Arbeitsalltag von Übersetzern ersichtlich wird. Die Ursache dafür liegt nicht allein in Zeitgründen:

> Reine Terminologiearbeit wird von vielen Auftraggebern nur in einem bestimmten Sinne gefordert. Meistens ist die Situation die folgende: Man beauftragt uns mit der Durchführung einer Übersetzung und stellt uns freundlicherweise zum Zwecke der Terminologieerfassung eine alte Übersetzung eines ähnlichen Texts zur Verfügung. Wir prüfen die Terminologie und stellen fest, daß sie schlecht bis abenteuerlich ist. Das sagen wir dem Auftraggeber und fragen, ob wir die Terminologie überarbeiten sollen, d. h. ob wir Terminologiearbeit leisten sollen. Die Antwort ist bedauerlicherweise sehr, sehr oft ein nachdrückliches Nein. Der Grund dafür ist nicht, daß die Leute etwas gegen Terminologiearbeit hätten, keineswegs; aber meistens der, daß das Handbuch, das wir übersetzen sollen, Teil einer umfangreichen Dokumentation ist, die ausnahmslos mit der von uns beanstandeten Terminologie übersetzt wurde und die man nicht bereit ist zu ändern. Und so verlangt man ausdrücklich von uns, daß wir uns darauf beschränken, die benötigte Terminologie aus der alten Übersetzung zu extrahieren bzw. zu erfassen und sie für den Einsatz bei der Neu-Übersetzung bereitzustellen[2].

2.2 Computerlinguistische Untersuchungen von Fachsprachen

2.2.1 Grundlagen der Forschung

Die neueren computerlinguistischen Ansätze zur Analyse von Fachsprachen und Fachtexten, insbesondere unter dem Aspekt der Übersetzung, sind durch zwei Entwicklungen beeinflußt:

- den Ergebnissen der Forschung zur maschinellen Übersetzung (MÜ)
- der Korpuslinguistik, einer neueren Richtung in der Linguistik und Computerlinguistik.

1. Konkrete Beispiele für diese Form der Terminologiearbeit finden sich bei Reinart 1993, Kap. 14.
2. Mossmann 1994, S. 26.

2.2.1.1 Maschinelle Übersetzung

Verfolgt man die Forschungsgeschichte der MÜ[1], so muß man feststellen, daß die hohen Erwartungen, die zu Beginn geäußert worden waren, immer mehr eingeschränkt oder revidiert wurden:

> Despite several decades of massive efforts, high-quality machine translation is still only possible in the case of some very restricted sublanguages ... in most cases it is more productive to support tools for human translators[2].

Es wird zwar weiterhin über MÜ geforscht, daneben enstand aber in den letzten Jahren eine Richtung, die sich mit der Interaktion zwischen menschlichen Übersetzern und elektronischen Hilfsmitteln im Übersetzungsprozeß beschäfigt: die computergestützte Übersetzung (*computer-aided translation: CAT*[3]). Diese Interaktion wird unter zwei Aspekten untersucht:

- CAT-C: Eine vom Menschen durch Präedition, Interaktion und/oder Postedition gesteuerte bzw. korrigierte Übersetzung durch den Computer
- CAT-H: Eine vorwiegend intellektuell angefertigte (Human-)Übersetzung.

  Hier geht es vor allem darum, die Arbeit des Übersetzers durch spezielle Hilfsmittel in Form von adäquaten Textverarbeitungswerkzeugen, elektronischen Wörterbüchern, Terminologiedatenbanken usw. möglichst ergonomisch zu unterstützen.

2.2.1.2 Korpuslinguistik

Texte bzw. Sammlungen von Texten (Korpora) waren schon immer ein Untersuchungsgegenstand sprachwissenschaftlicher und computerlinguistischer Forschungen. Korpusbasierte Studien in größerem Umfang stellen jedoch einen neuen Trend in der Computerlinguistik dar. Das zeigt sich z. B. in zahlreichen Kongressen oder in den zwei Sonderheften der Fachzeitschrift *Computational Linguistics* zu diesem Thema. Diese Situation ist auf verschiedene Faktoren zurückzuführen, die einerseits Ausdruck methodologischer Überlegungen sind, und andererseits technologische Fortschritte widerspiegeln.

Die formale Linguistik hatte den Anspruch vertreten, Sprache (in Form beliebiger Texte) formal beschreiben zu können. Die Versuche, die sprachliche Komplexität mit Hilfe von regelbasierten Methoden, die u. a. von den Arbeiten Chomskys inspiriert sind, in den Griff zu bekommen, haben nur teilweise zum Erfolg geführt. Die Diskussion um die verschiedenen Formalismen hatte zur Folge, daß immer weniger Phänomene immer eleganter erklärt werden konnten und die Beispiele für Außenstehende

1. Zur Geschichte der MÜ-Forschung vgl. Schwanke 1991, die vor allem die Situation in der USA beschreibt.
2. Isabelle 1992, S. 1.
3. Frz.: Traduction assistée par ordinateur (TAO).

immer weniger nachvollziehbar wurden. Der Versuch, Grammatiken mit größerem Abdeckungsgrad zu erstellen, hat zu dem Eindruck geführt, daß die Aufgabe langwierig, mit den gegenwärtigen Formalismen nicht zu machen und prinzipiell endlos sei. Diese “Erkenntnis” führte zu einer Wiederentdeckung von Ideen, die in den 50er Jahren als empirische bzw. statistische Modelle bekannt geworden waren. Zu dieser Zeit wurde z. B. die maschinelle Übersetzung als ein reines Dekodierproblem gesehen, doch waren die Computerresourcen damals bei weitem nich adäquat für eine Bearbeitung der Daten gemäß diesem Modell. Der Fortschritt in der Technologie und Leistungsfähigkeit von Computern hat die Wiedereinführung und Neubelebung dieses Ansatzes gefördert. In gleichem Maße dazu beigetragen hat die **zunehmende Verfügbarkeit von großen textuellen Resourcen in maschinenlesbarer Form**, an denen statistische Methoden gezielt angewandt und getestet werden können.

Statistische Methoden werden für die Lösung verschiedener linguistischer Fragestellungen eingesetzt, z. B. in der Spracherkennung, bei der automatischen Wortklassenannotation (*POS-Tagging*), beim Parsing (präferenzbasierte Parser) und in der monolingualen und bilingualen Lexikographie. Eine Vorreiterrolle haben statistische Sprachmodelle in der Spracherkennung gespielt, mit denen im praktischen Einsatz klare Dominanz über regelbasierte Ansätze erzielt werden konnte.

2.2.1.2.1 Untersuchungen multilingualer Textkorpora

Die schon erwähnte zunehmende Verfügbarkeit textueller Resourcen in elektronischer Form führte zu einer Reihe von korpuslinguistischen Arbeiten. Unter den verfügbaren Resourcen finden sich auch immer mehr multilinguale Texte, d. h. Versionen eines Textes in mehreren Sprachen, die eine wertvolle Quelle für die Erstellung von Hilfsmitteln für den Übersetzungsprozeß sind.

Die *Canadian Hansards* (die engl. und frz. Versionen der kanadischen Parlamentsdebatten) bzw. deren Bereitstellung in elektronischer Form waren der Ausgangspunkt einiger Studien[1], die vor allem in den USA durchgeführt wurden. Korpuslinguistische Verfahren, die für die Analyse einsprachiger Texte entwickelt worden waren (z. B. die Bestimmung lexikalischer Korrelationen durch Berechnung der *Mutual Information*) konnten so auch für die Analyse **bilingualer Texte** getestet werden. Darüber hinaus wurden spezielle Verfahren der Textanalyse entwickelt, die die Tatsache nutzen, daß bilinguale Texte übersetzungen voneinander sind. Diese Verfahren sind nicht nur unter dem Aspekt der Übersetzung interessant, sondern liefern auch Ergebnisse, die für einzelsprachliche Studien relevant sein können (z. B. bei der Disambiguierung von Lesarten). Im Mittelpunkt der meisten korpuslinguistischen Arbeiten zu bilingualen Texten steht jedoch die Entwicklung von Verfahren, die eine automatische Zuordnung von Textsegmenten eines Textes mit Textsegmenten des anderen Textes[2] (**Alignierung**) ermöglichen.

1. Kupiec 1993, Klavans & Tzoukermann 1990.
2. Ein Text ist die Übersetzung des anderen Texts.

Einige dieser Arbeiten sollen hier vorgestellt werden. Sie wurden unter folgenden Gesichtspunkten ausgewählt:

- Sie spiegeln in etwa den Stand der Technik auf diesem Gebiet wider.
- Sie wurden für Anwendungen im “realen” Übersetzungskontext konzipiert.
- Sie beziehen sich auf fachsprachliche Texte.
- Sie beziehen sich auf die Sprachen, in denen auch das Korpus des Europäischen Patentamts, das Grundlage für die vorliegende Arbeit[1] ist, vorliegt.

2.2.2 Das Programm Termight

Termight[2] ist ein Programm, das Übersetzer und Terminologen bei der Erstellung von Glossaren unterstützen soll. Ein Glossar ist eine Liste von Fachtermini und ihrer Übersetzungen. Die Erstellung eines Glossars gliedert sich in zwei Teilaufgaben:

- Generierung einer Liste von Termini (monolinguale Aufgabe)
- Bestimmung der Übersetzungsäquivalente (bilinguale Aufgabe).

Potentielle Fachtermini und ihre Übersetzungen werden aus bilingualen Texten mit Hilfe einer automatischen Wortklassenzuordnung (*POS-Tagging*) und einem Wortalignierungsalgorithmus extrahiert. Der Extraktionsalgorithmus wurde so konzipiert, daß möglichst alle potentiellen Fachtermini eines Textes lokalisiert werden, um zu vermeiden, daß wichtige, aber weniger häufige Termini bzw. deren Übersetzungen nicht erfaßt werden. Da dieser Ansatz jedoch zu einer höheren Fehlerrate der extrahierten Einheiten führen kann, wurde von den Autoren eine Phase der Filterung vorgesehen, die vom Benutzer (Übersetzer oder Terminologe) durchgeführt wird. Termight schließt die Darstellung der potentiellen Fachtermini in einer möglichst ergonomischen Art und Weise ein (Strukturierung der Fachtermini, Konkordanzen). Das Programm wird von einem größeren Übersetzungsdienst[3]genutzt und evaluiert.

2.2.2.1 Monolinguale Aufgabe

Die Autoren gehen davon aus, daß die meisten Fachtermini Nominalphrasen sind, die aus mehreren Wörtern bestehen und nach einigen wenigen Typen von Bildungsmustern aufgebaut sind. Termight extrahiert jedoch sowohl mehrgliedrige Nominalphrasen als auch einzelne Wörter.

1. In diesem Zusammenhang ist die Untersuchung des engl.-niederländischen Korpus von Eijk (siehe Abschnitt 2.2.4 Der Ansatz von Eijk, Seite 17) interessant, in der Besonderheiten der niederländischen Wortbildung angesprochen werden, die viele Gemeinsamkeiten mit der dt. Wortbildung hat.
2. Church & Dagan 1994.
3. AT&T Business Translation Services.

Dagan und Church berichten über Anwendungstests mit bilingualen engl.-frz. und engl.-dt. Texten. Das Extraktionsprogramm wird in dem Artikel nur kurz beschrieben. Die Angaben zur Bestimmung der potentiellen Termini und zu ihrer Aufbereitung für den Benutzer beziehen sich nur auf die engl. Texte. Die Autoren beschreiben die aus den engl. Texten extrahierten Einheiten folgendermaßen:

> The list of candidate terms contain both multi-word noun phrases and single words. The multiword terms match a small set of syntactic patterns defined by regular expressions and found by searching a version of the document tagged with parts of speech. The set of syntactic patterns is considered as a parameter and can be adapted to a specific domain by the user. Currently our patterns match only sequences of nouns, which seem to yield the best hit rate in our environment. Single-word candidates are defined by taking the list of all words that occur in the document and do not appear in a standard stop-list of "noise" words[1].

Die potentiellen Fachtermini werden so sortiert, daß alle Nominalphrasen, die denselben "Kopf" (*head*) haben, in Gruppen zusammengefaßt werden. Die Gruppen werden nach absteigender Frequenz der Headwords angeordnet. Diese Frequenz steht i. a. in Korrelation zu der Wahrscheinlichkeit, daß das entsprechende Headword in einem Fachterminus verwendet wird. Die zu einem Headword gehörenden Termini werden in einer Form sortiert, die der Reihenfolge der Modifizierung in einfachen engl. Nominalphrasen entspricht. Die Sortierung führt solche Termini zusammen, die verschiedene Modifikationen eines allgemeineren Terminus darstellen, z. B. die Termini *default paper size*, *paper size* und *size*.

Die potentiellen Termini werden in der beschriebenen Anordnung zusammen mit Konkordanzen (in diesem Fall mit den Sätzen, in denen die potentiellen Termini vorkommen) dem Benutzer zur Beurteilung vorgelegt.

Evaluierung

Es wird über einen Test berichtet, in dem aus einem engl. Text von 300 000 Wörtern eine Liste von 1 700 potentiellen Termini in 10 Stunden extrahiert wurde. Die Liste wurde von einem Terminologen mit einer Geschwindigkeit von 150-200 Termini pro Stunde überprüft. Das entsprach etwa der Hälfte der Zeit, die der Terminologe zur Ausführung dieser Aufgabe bräuchte, wenn er nur auf einfachere lexikographische Hilfsmittel zurückgreifen könnte.

2.2.2.2 Bilinguale Aufgabe

In den meisten Programmen, mit denen Übersetzungsäquivalenzen von Fachtermini aus bilingualen Texten bestimmt werden, erfolgt die bilinguale Zuordnung der Fachtermini auf der Basis einer Alignierung auf der Satzebene (siehe Kapitel 4 Satzalignie-

1. Church & Dagan 1994, S. 35.

rung, Seite 31). Im Gegensatz dazu wird in Termight ein Modul (*Word_align*) verwendet, das die Texte auf der Wortebene aligniert[1]. Die Autoren stellten bei Testläufen fest, daß nur eine Alignierung auf der Wortebene auch die Lokalisierung von Entsprechungen für seltenere Termini ermöglicht. In den satzbasierten Verfahren erfüllen solche selteneren Termini die statistischen Signifikanzkriterien in den meisten Fällen nicht. Wenn zumindest ein Teil der Wörter, aus denen ein seltener Terminus besteht, mehrmals im Text vorkommt, so können diese Wörter durch das Modul *Word_align* korrekt aligniert werden. In diesem Fall kann Termight zumindest eine Übersetzung lokalisieren, die sich mit der korrekten überlappt.

Für jedes Vorkommen eines Terminus im Quelltext lokalisiert Termight ein mögliches Übersetzungsäquivalent. Dieses wird durch die Sequenz von Wörtern definiert, die zwischen der ersten und der letzten Position der Wörter im Zieltext stehen, mit denen ein Wort des quellsprachlichen Terminus aligniert ist.

> Beispiel: Der Terminus *optional parameters box* aus dem engl. Quelltext wird mit der Sequenz *zone paramètres optionnels* aus dem frz. Zieltext aligniert, weil die frz. Wörter *zone* und *optionnels* zuvor mit Bestandteilen des englischen Terminus aligniert worden waren.

Für alle Vorkommen eines quellsprachlichen Terminus werden auf diese Art Kandidaten für Übersetzungsäquivalenzen bestimmt. Sie werden in einer Liste nach absteigender Frequenz angeordnet und zusammen mit den jeweiligen Konkordanzen dem Benutzer zur Validierung vorgelegt.

Evaluierung

Die Performanz der bilingualen Komponente von Termight wurde mit 192 engl. Termini aus den engl. und dt. Versionen eines technischen Handbuchs untersucht. In allen Fällen befand sich die korrekte Übersetzung in der Liste der möglichen Äquivalente. Sie war der erste “Kandidat” in 40 % der Fälle, der zweite in 7 % der Fälle.

2.2.3 Der Ansatz von Gaussier

Gaussiers Arbeit[2] hat zum Ziel, mit statistischen Methoden aus einem bilingualen frz.-engl. Korpus Entsprechungen von lexikalischen Einheiten zu extrahieren. Die lexikalischen Einheiten sind allgemeinsprachlicher oder fachsprachlicher Natur und können sowohl Einzelwörter als auch Komposita sein. Die Arbeit ist in ein größeres Forschungsprojekt integriert und baut u. a. auf den Untersuchungen von B. Daille (1995)[3] auf, in der frz. und engl. Komposita aus bilingualen engl. und frz. Fachtexten extrahiert und mit statistischen Methoden bewertet wurden.

1. Dagan et al. 1994.
2. Gaussier 1995.
3. Vgl. dazu auch Abschnitt 5.2.3 Ansatz von Daille, Seite 72.

Ein bilinguales Korpus aus dem Bereich der Telekommunikation (200 000 Wörter pro Sprache) wurde auf Satzebene aligniert[1]. Die Beobachtung, daß zwei Wörter, die die Übersetzung voneinander sind, in den meisten Fällen ähnliche Distributionen aufweisen, ist der Ausgangspunkt für die Entwicklung eines Algorithmus zum Auffinden der Entsprechungen. Dies erfordert ein Berechnungsmaß, mit dem Korrelationen zwischen den Distributionen ermittelt werden können. In der Statistik gibt es mehrere Modelle zur Berechnung von Korrelationen, unter denen Gaussier zwei auswählt. Dies ist zum einen die *Mutual Information*, die für linguistische Problemstellungen häufig verwendet wird[2], zum anderen der Ähnlichkeitsquotient nach Dunning (1993), der den Vorteil hat, daß damit sowohl häufige als auch seltene Phänomene angemessen bewertet werden können. Daille hat in ihrer Arbeit eine ganze Reihe statistischer Modelle zur Berechnung von Korrelationen evaluiert und kommt zu dem Ergebnis, daß die meisten Modelle seltene Phänomene überbewerten.

Gaussier ergänzt die verwendeten Modelle durch einen neuen Parameter (*distorsion*), der die Position der Wörter innerhalb des Satzes ausdrückt. Er geht davon aus, daß der Position eines Wortes im quellsprachlichen Satz ein Ausschnitt (“Fenster”) im zielsprachlichen Satz entspricht, in dem die Übersetzung des quellsprachlichen Wortes sich mit größter Wahrscheinlichkeit befindet. Die optimale Größe dieses Ausschnitts wurde empirisch bestimmt.

2.2.3.1 Alignierung einzelner Simplizia

Aus dem satzweise alignierten, mit Wortklassen annotierten und auf die Grundformen lemmatisierten Korpus wurde das Vokabular in Form einzelner Wörter extrahiert. Wörter, die zu den geschlossenen Wortklassen gehören (z. B. Pronomen, Artikel usw.) und Wörter bzw. Grundformen, deren Auftretenshäufigkeit unter 2 liegt, wurden in der Alignierung nicht berücksichtigt. Für die Evaluierung wurde manuell eine Referenzliste erstellt, die alle im Text vorkommenden Übersetzungen eines Wortes enthält. Für jedes Wortpaar, bestehend aus Wörtern des Quelltexts und des Zieltexts, die in miteinander alignierten Sätzen vorkommen, wurden drei Berechnungen durchgeführt: *Mutual Information* mit Berücksichtigung der Positionen der Wörter in den Sätzen, *Mutual Information* ohne Berücksichtigung der Positionen in den Sätzen, und Ähnlichkeitsquotient nach Dunning. Die Wortpaare bestehen immer aus Wörtern derselben Wortklasse. Die Liste der ermittelten Entsprechungen wurde manuell evaluiert. Unter den ersten 40 Entsprechungen, d. h. den Entsprechungen mit den höchsten Korrelationswerten, lag die Genauigkeitsquote[3] (*precision*) zwischen 47 und 55%, die Vollständigkeitsquote (*recall*) zwischen 84 und 99 %. Die besten Resultate für die Genauigkeits- und die Vollständigkeitsquote wurden mit dem Ähnlichkeitsquotienten erzielt. Die Fehler sind auf inkorrekte Satzalignierung, Zuordnung einer falschen

1. Beschreibung des Algorithmus siehe Abschnitt 4.3.5 Erweiterung der Church-Gale-Methode, Seite 48.

2. Z. B. bei Church & Hanks 1990, Calzolari & Bindi 1990.

3. Übersetzungen für *recall* und *precision* aus Henzler 1992, S. 165.

Wortklasse und auf die Tasache, daß einige Wörter keine gleichbleibende Übersetzung haben, zurückzuführen.

2.2.3.2 Alignierung von Komposita

Die linguistische Beschreibung und die Extraktion der engl. und frz. Komposita baut auf den Untersuchungen von Daille (1995) auf. Komposita werden durch ihr Bildungsmuster, dargestellt als Abfolge von Wortklassen, definiert. Für das Frz. werden Komposita der Form[1] “Nomen *de* Nomen”, “Nomen Adjektiv und “Nomen Präposition[2] Nomen” berücksichtigt, für das Engl. Komposita der Form[3] “Nomen Nomen” und “Adjektiv Nomen”. Sequenzen, die diesen Bildungsmustern entsprechen, wurden mit Automaten aus den Texten extrahiert. Unter diesen Sequenzen befinden sich auch Nominalphrasen, die keine Komposita sind[4]. Ein großer Teil der Komposita sind laut Gaussier Fachtermini.

Die Alignierung der Komposita erfolgte nach zwei Methoden, einerseits über die Alignierung der Wörter, aus denen die Komposita bestehen, andererseits durch eine direkte Alignierung der Komposita.

In beiden Fälle wurden nur Komposita mit einer Auftretenshäufigkeit größer 3 aligniert. So konnten Fehler, die auf die inkorrekte Extraktion der Komposita zurückzuführen sind, möglichst gering gehalten werden. Bei der zweiten Methode werden auch die Wahrscheinlichkeiten von Strukturäquivalenzen berücksichtigt. Es lassen sich Tendenzen beobachten, beispielsweise daß die engl. Übersetzung eines frz. Kompositums der Struktur “Nomen *de* Nomen” in 80 % der Fälle einer Struktur der Form “Nomen Nomen” entspricht[5]. Für die Implementierung der Methoden wurden außer der *Mutual Information* und dem Ähnlichkeitsquotienten weitere Korrelationsmodelle aus der Statistik verwendet. Eine Kombination der beiden Methoden - basierend auf einzelnen Wörtern und auf Komposita - zeigte die besten Ergebnisse: 80% der ersten 700 Komposita-Alignierungen waren korrekt.

2.2.4 Der Ansatz von Eijk

Das Ziel der Arbeit[6] besteht darin, die Generierung von bilingualen Listen von Fachtermini aus Texten und deren Übersetzungen so weit wie möglich zu automatisieren. Eine bilinguale Liste von Fachtermini ist eine Liste quellsprachlicher Termini, wobei jeder Terminus mit einer Rangliste möglicher zielsprachlicher Übersetzungsäquiva-

1. Gaussier 1995 (Kap. 4.1.1.), siehe dazu auch Abschnitt 5.2.3 Ansatz von Daille, Seite 72.
2. D. h. andere Präpositionen als “de”.
3. Gaussier 1995, Kap. 4.1.2.
4. Diese Problem wird von Daille 1995 ausführlich erörtert.
5. Maxwell 1992.
6. Eijk 1993.

lente versehen ist. Das Testkorpus ist die engl. und niederländische Version der Darstellung eines europäischen Forschungsprojekts (ca. 25 000 Wörter pro Sprache).

Die Erstellung der Liste besteht aus zwei Teilaufgaben:

- Extraktion der Termini für jede Sprache (mit linguistischen Methoden)
- Bestimmung von Übersetzungsäquivalenzen (mit statistischen Methoden).

Vorbereitung der Texte und Extraktion

Die Texte wurden auf Satzebene aligniert[1] und automatisch mit Wortklassen annotiert. Die Extraktion der potentiellen Fachtermini erfolgte durch partielles Parsing mit einem Pattern-Matching-Verfahren.

2.2.4.1 Statistische Bestimmung der Übersetzungsäquivalente

Zur Bewertung der Wahrscheinlichkeit der Entsprechung zwischen einem quellsprachlichen und einem zielsprachlichen Terminus wird ein Entsprechungsmaß verwendet, das auf lokalen und globalen Frequenzen der Termini basiert:

> This measure is based on the intuition that the translation of a term is likely to be more frequent in the subset of target text segments aligned to source text segments containing the source language term than in the entire target language text[2].

Die *globale Frequenz* ist die Auftretenshäufigkeit eines Terminus im gesamten Text. Die *lokale Frequenz* bezieht sich auf die Auftretenshäufigkeit in den Textsegmenten, die miteinander aligniert sind und in denen der quell- bzw. zielsprachliche Terminus, die möglicherweise Übersetzungsäquivalente sind, auftreten. Da dieses Maß seltene Termini überbewertet, wurde ein Schwellenwert als Filter definiert.

2.2.4.2 Ergebnisse

Eijk testet verschiedene Methoden zur Berechnung des Entsprechungsmaßes anhand von 1190 niederländischen Nominalphrasen. Das beste Ergebnis lag bei einer Vollständigkeitsquote von 62 % und einer Genauigkeitsquote von 88 %. Berechnungen, die auf den Frequenzen der Nominalphrasen basieren, zeigten bessere Ergebnisse als Berechnungen, die auf den Frequenzen der Wörter basieren, aus denen diese Nominalphrasen bestehen. Das hängt auch damit zusammen, daß im Niederländischen (ähnlich wie im Deutschen) die Kompositabildung von Nomina zu einer orthographisch zusammenhängenden Form führt. Im Gegensatz zum Englischen, wo die Bestandteile eines Kompositums meist noch orthographisch (und damit automatisch) erkennbar bleiben.

1. Church-Gale-Methode, siehe Abschnitt 4.3 Methode von Church und Gale, Seite 43.
2. Eijk 1993, S. 115.

Die Fehler in der Vollständigkeitsquote sind auf Fehler in der Vorverarbeitung (85 %) und auf einzelsprachliche Besonderheiten (15 %) zurückzuführen:

- inkorrekte Satzalignierung (6%)
- inkorrekte Extraktion aufgrund fehlerhafter Wortklassenannotation (15 %)
- inkorrekte Extraktion aufgrund fehlender Erkennung von adverbialen und präpositionalen Mehrwortverbindungen (6 %), z. B.wird aus der engl. Sequenz *with respect to* der “Terminus” *respect* extrahiert
- Die syntaktischen Strukturen der Termini in den beiden Sprachen entsprechen sich nicht (47 %). Das Pattern-Matching-Verfahren erkennt in beiden Sprachen keine postnominalen präpositionalen Phrasen. Eijk stellt fest, daß aber oft eine postnominale Phrase im Niederländischen dem ersten Bestandteil eines engl. Kompositums entspricht.
- Nominalphrasen werden nicht durch Nominalphrasen übersetzt (15 %), sondern durch Verbalphrasen, Adverbien usw.

Eijk betont, daß diese Ergebnisse an einem größeren Textkorpus überprüft werden müssen. Er geht davon aus, daß dadurch vor allem die statistischen Verfahren zur Bestimmung der Übersetzungsäquivalente erheblich verfeinert und verbessert werden könnten. Es ist aber unklar, ob er sich davon auch eine bessere Extraktion der Termini verspricht, die aber die Voraussetzung für eine Verbesserung des Gesamtergebnisses ist.

2.2.5 Das EURAMIS-Projekt der Europäischen Gemeinschaft

Die Europäische Gemeinschaft unterhält den größten Übersetzungsdienst der Welt[1], der mit einer Reihe von Hilfsmitteln verschiedenster Art ausgestattet ist.

Das Projekt EURAMIS[2] (“European Advanced Multilingual Information System” verfolgt zwei Zielsetzungen:

- Die Bereitstellung eines Interface, das den Zugang und die Benutzung verschiedener schon existierender Dienstleistungen und Hilfsmittel (z. B. Maschinelle Übersetzung, juristische Datenbanken, Terminologiedatenbanken usw.) organisiert.

  Diese Aufgabe ist zuerst ein Soft- und Hardwareproblem. Zum einen sind manche Applikationen, vor allem, wenn sie sehr umfangreich sind, auf dem PC nicht lauffähig. Zum anderen wird es auch in absehbarer Zeit nicht möglich sein, alle von einem linguistischen Standpunkt her nützlichen Daten auf einem PC zu halten; beispielsweise enthält die Terminologiedatenbank EURODICATOM allein schon mehr als 600 000 multilinguale Einträge.

1. Pro Jahr werden mehr als eine Million Seiten an Übersetzungen angefertigt. Eine Beschreibung der Arbeitsabläufe findet sich bei King 1995.

2. Beschrieben u. a. in Blatt 1995.

- mittelfristig die Aufbereitung existierender linguistischer Resourcen (z. B. paralleler Texte). Die Resourcen sollen unter verschiedenen Aspekten strukturiert und wiederauffindbar gemacht werden, so daß sie auf Benutzeranfragen zur Verfügung gestellt werden können.

2.2.5.1 Datenbank der linguistischen Resourcen

Das Kernstück aller EURAMIS-Applikationen ist eine Datenbank linguistischer Resourcen (*Linguistic Resources Database*: LRD). Sie wird genutzt, um alle für die verschiedenen Anwendungen notwendigen Daten zu speichern, zu strukturieren und aufzubereiten. Die Daten sind beispielsweise Originaltexte und ihre Übersetzungen, Terminologie aus verschiedenen Datenbanken, Glossare sowie Lexika aus MÜ-Systemen. Es ist geplant, zu einem späteren Zeitpunkt auch allgemeinsprachliche elektronische Wörterbücher zu integrieren. Eine flexible Gestaltung der Datenbank trägt der unterschiedlichen Natur der Einträge Rechnung.

Ein wichtiger Bestandteil der LRD ist die Sammlung multilingualer, satzweise alignierter Texte (*multilingually aligned texts*: *MATs*). Diese Texte sind mit einzelsprachlichen Beschreibungen der in den Texten enthaltenen linguistischen Einheiten (*linguistic objects*: *LOs*), z. B. einzelner Wörter, komplexer Phrasen, Sätze und Absätze annotiert.

Verschiedene Versionen der Datenbank bestehen nebeneinander: eine zentrale Datenbank, die für alle Benutzer zugänglich ist und nur von wenigen autorisierten Personen verändert werden kann, Datenbanken mit Zugangsberechtigungen für Arbeitsgruppen (wie z. B. Institutionen, Direktorate, Gruppen zu einzelnen Themenbereichen, Übersetzungsabteilungen) und “private” Datenbanken zur individuellen Benutzung.

2.2.5.2 Anwendungen

Unter den teils schon realisierten und teils erst geplanten Anwendungen sollen hier nur die wichtigsten erwähnt werden[1].

Übersetzungsspeicher

Der Übersetzungsspeicher (engl. *translation memory*) bzw. die dort integrierten Programme erfüllen drei Aufgaben:

- maschinelle Aufbereitung existierender Daten (z. B. Alignierung von Texten und deren Übersetzung auf Satzebene)
- Analyse neuer Daten (z. B. wird für die Anfertigung einer neuen Übersetzung soviel Information wie möglich aus bestehenden Übersetzungen bereitgestellt).

1. Vgl. dazu Blatt 1995 und King 1995.

- Präsentation und Edition der Ergebnisse unter inhaltlichen und ergonomischen Gesichtspunkten.

Der in EURAMIS verfolgte Ansatz unterscheidet sich in wesentlichen Punkten von den Übersetzungsspeichern, die zur Zeit auf dem Markt erhältlich sind.

Die herkömmlichen Alignierungstechniken basieren in erster Linie auf Formatierungsinformation und Satzlängen (vgl. dazu Kapitel 4 Satzalignierung, Seite 31). Im Gegensatz dazu werden in EURAMIS zur Satzalignierung auch Informationen aus der LRD (z. B. Entsprechungen zwischen einzelnen Wörtern) herangezogen.

Der Übersetzungsspeicher kann auch für die Ersetzung von Textteilen verwendet werden. Dazu gibt der Benutzer existierende Lexika, Glossare bzw. eigene Listen an, auf denen die Ersetzung basieren soll. Es kann definiert werden, ob die Entsprechungen zwischen den Wörtern des Textes und den Einträgen in den Lexika und Glossaren vollständig sein muß oder auch partiell sein kann (*fuzzy matches*).

Programme für die Erstellung neuer Dokumentversionen

Diese Programme nutzen die Informationen aus dem Übersetzungsspeicher. Da jedes Dokument in der LRD eine Referenz auf das quellsprachliche Dokument enthält, aus dem es übersetzt wurde, können dem Benutzer Dokumente angeboten werden, die identische oder annähernd identische Teile (Bestimmung durch statistische Methoden) enthalten. Die Unterschiede zwischen dem neuen und dem alten Dokument werden entsprechend graphisch markiert.

Terminologie

Die schon bestehende terminologische Datenbank EURODICATOM wird im Rahmen von EURAMIS zur Verfügung gestellt und mit besseren Suchmöglichkeiten (Wildcards, Boolesche Operatoren, Präferenzen usw.) ausgestattet. Weitere interne oder externe Datenbanken sollen ebenfalls in EURAMIS integriert werden.

Es sind Möglichkeiten für die Erstellung neuer Glossare und deren Bereitstellung für eine große Benutzergruppe vorgesehen. Die automatische Generierung von textbezogenen Glossaren ist bisher noch auf die Sprachen beschränkt, die von den MÜ-Systemen der Europäischen Gemeinschaft abgedeckt werden. Auch hier sind Erweiterungen vorgesehen.

2.2.6 Fazit

Die folgenden Ausführungen beziehen sich vor allem auf die drei computerlinguistischen Arbeiten, die vorgestellt wurden. Das Projekt EURAMIS verfolgt einen praxisorientierten Ansatz, in dem es primär um die Nutzbarmachung vorhandener Resourcen geht, deren Bereitstellung zuerst die Lösung nicht zu unterschätzender soft- und hardwaretechnischer Fragen erfordert.

Die Arbeiten von Church und Dagan, Gaussier und Eijk unterscheiden sich in ihrer Zielsetzung kaum und basieren auf sehr ähnlichen Ansätzen. Fachtermini werden als Nominalphrasen definiert, die nach bestimmten Bildungsmustern aufgebaut sind. Genauere Beschreibungen dieser Bildungsmuster fehlen meist. Die Nominalphrasen werden aus den annotierten Korpora in einem Pattern-Matching-Verfahren extrahiert. Fehler in der Extraktion werden durch die Festsetzung von Schwellenwerten für die Auftretenshäufigkeit und andere statistische Filter kompensiert. Man geht davon aus, daß seltenere Phänomene eher inkorrekt sind als häufig auftretende. Das hat zur Folge, daß seltenere Nominalphrasen für die weiteren Schritte der Analyse nicht mehr zur Verfügung stehen. Das Programm Termight stellt hier eine Ausnahme dar: alle Nominalphrasen werden in einem interaktiven Verfahren von einem Experten ausgewertet und werden, falls sie korrekt sind, weiterverarbeitet. Vor einer solchen Auswertung haben im Prinzip alle extrahierten Nominalphrasen den Status *potentieller* Termini. Korrekterweise müßte also immer von potentiellen Fachtermini und nicht nur von Fachtermini die Rede sein. Dieser Aspekt wird aber in den Arbeiten von Gaussier und Eijk nicht problematisiert.

Die Bestimmung der Übersetzungsäquivalente der Fachtermini liegt die (implizite) Hypothese zugrunde, daß sich die Fachtermini in den beiden untersuchten Sprachen strukturell entsprechen: Es wird angenommen, daß Nominalphrasen immer durch Nominalphrasen übersetzt werden und darüber hinaus wird davon ausgegangen, daß Nominalphrasen einer begrenzten Anzahl von Bildungsmustern in der einen Sprache auch Nominalphrasen einer begrenzen Anzahl von Bildungsmustern in der anderen Sprache entsprechen. Eijk zeigt, daß diese Hypothese zumindest für das Sprachpaar Englisch-Niederländisch bei weitem nicht immer gültig ist.

Die Korrektheit der ermittelten Übersetzungsäquivalenzen werden bei Gaussier und Eijk durch statistische Methoden überprüft, wobei versucht wird, mit Schwellenwerten oder anderen Berechnungen aus allen Entsprechungen die herauszufiltern, die mit größter Wahrscheinlichkeit richtig sind. Bei der Konzeption von Termight wurde eine andere Strategie verfolgt: die ermittelten Entsprechungen werden von einem Experten überprüft, der bei seiner Arbeit durch zusätzliche Hilfsmittel wie z. B. Konkordanzen unterstützt wird.

2.3 Entwicklung eines eigenen Verfahrens

Das Verfahren soll sowohl linguistisch adäquat sein als auch so viel Arbeitsschritte wie möglich automatisieren. Ein vollautomatisches Verfahren wird nicht angestrebt, weil damit i. a. nur für einen geringen Teil der potentiellen Fachtermini zuverlässige Aussagen gemacht werden können und weil ich mir außerdem nicht vorstellen kann, daß sich ein Übersetzer bei der Lösung terminologischer Probleme auf ein vollautomatisch erstelltes Glossar verlassen würde.

In der Übersetzungswissenschaft werden Fachtermini vor allem über semantische Kriterien definiert. Solche Kriterien sind jedoch bei der Konzeption eines (semi-)automa-

tischen Verfahrens nicht operationalisierbar. Andererseits haben die korpuslinguistischen Ansätze trotz aller Mängel durchaus zu Ergebnissen geführt, die für die praktische Terminologiearbeit von Nutzen sind. Church und Dagan, Eijk und Gaussier arbeiten jedoch mit einer Reihe von linguistischen Hypothesen über Termini und deren Übersetzungsäquivalente, deren Gültigkeit meiner Ansicht nach genauer überprüft werden sollte.

Vor diesem Hintergrund erschien es mir sinnvoll, ein Verfahren zur Terminologieextraktion in folgende Arbeitsschritte zu gliedern:

- Bereitstellung eines Korpus fachsprachlicher Texte, die möglichst konsistent und einheitlich übersetzt wurden
- Vorverarbeitung des Korpus (Segmentierung der Texte in kleinere Einheiten und Satzendeerkennung)
- Synchronisierung des Korpus auf Satzebene (Satzalignierung)
- Linguistische Beschreibung der Fachtermini für jede Einzelsprache

  Für ein automatisches Verfahren kommen nur Ansätze in Frage, in denen Fachtermini mit **morphosyntaktischen Kriterien** definiert werden. Dies erfordert eine entsprechende maschinelle Aufbereitung der Texte.
- Suche nach Werkzeugen für Lemmatisierung und Zuordnung von Wortklassen

  Es ist zu prüfen, welchen Korrektheitsgrad und welche Abdeckung die automatische Aufbereitung der Texte mit diesen Hilfsmittel ermöglicht; es muß nach Lösungen für die Behebung eventueller Fehlerquellen gesucht werden.
- Implementierung eines Verfahrens zur Extraktion potentieller Fachtermini
- Evaluierung

  Die extrahierten Sequenzen sollen unter einzelsprachlichen und kontrastiven Aspekten evaluiert werden.

  Es ist für jede Einzelsprache zu untersuchen, ob die extrahierten Sequenzen linguistisch korrekt sind, um dann in einem zweiten Schritt zu prüfen, inwieweit es sich dabei um Fachtermini handelt. In der Literatur finden sich zwar viele Beschreibungen darüber, wie statistische Filtermethoden zur Validierung potentieller Fachtermini eingesetzt werden, die Aussagen darüber sind jedoch sehr widersprüchlich. Dieses Problem kann, wie bereits erwähnt, hier nicht gelöst werden.)

  Anhand einer kontrastiven Untersuchung soll überprüft werden, inwieweit sich die für jede Einzelsprache extrahierten potentiellen Fachtermini inhaltlich entsprechen und als Übersetzungsäquivalente betrachtet werden können.

3 Das Korpus

Die Besonderheiten des Korpus und die Auswahlkriterien für die Teilkorpora müssen im Zusammenhang mit den in der Patentdokumentation vorkommenden Textsorten gesehen werden. Deshalb werden zuerst in einem Exkurs einige Grundsätze des europäischen Patentwesens erläutert.

Es folgt eine Beschreibung des Korpus des Europäischen Patentamts (EPA-Korpus), der für die vorliegende Untersuchung zur Verfügung stand und der aus zwei Korpora besteht. Dann wird die Strukturierung der einzelnen Dokumente sowie die Kriterien, die für die Erstellung von Teilkorpora genutzt werden konnten, erläutert. Abschließend sollen zwei Arbeiten vorgestellt werden, in denen Patenttexte unter syntaktischen und textlinguistischen Aspekten untersucht wurden.

3.1 Exkurs: Das Europäische Patentwesen

Was ist ein Patent?

> Ein Patent schützt eine Erfindung. Der Patentinhaber ist befugt, Dritten die Nutzung seiner Erfindung für einen bestimmten Zeitraum in einem bestimmten Land zu untersagen. Ohne seine Zustimmung darf niemand die geschützte Erfindung benutzen, d. h. sie darf von Dritten weder hergestellt noch angeboten, noch in Verkehr gebracht noch sonstwie zu gewerblichen Zwecken benutzt werden. Das Patent verbietet auch Importe von geschützten Produkten aus Ländern, in denen die Erfindung nicht patentiert ist.

Was ist patentierbar?

> Eine patentgeschützte Erfindung ist ein Erzeugnis, eine Vorrichtung, ein chemischer Stoff, ein Verfahren oder eine Verwendung. Eine Erfindung ist ein Vorschlag für die praktische Verwirklichung einer Idee zur erfolgreichen Lösung eines technischen Problems.

Patente werden nur für Erfindungen erteilt, die

- neu sind
- auf einer erfinderischen Tätigkeit beruhen und
- gewerblich anwendbar sind.

> Die Erfindung ist **neu**, wenn sie zum Zeitpunkt der Anmeldung nicht in irgendeiner Form (schriftlich, mündlich oder aus der Praxis) bekannt war, d. h. nicht zum “Stand der Technik” gehörte. Sie beruht auf einer **erfinderischen Tätigkeit**, wenn sie sich für den Fachmann nicht in naheliegender Weise aus dem Stand der Technik ergibt. Sie ist **gewerblich anwendbar**, wenn sie auf irgendeinem gewerblichen Gebiet (inklusive der Landwirtschaft) hergestellt oder benutzt werden kann.

Europäisches Patent

Mit einem europäischen Patent kann der Anmelder auf Antrag Patentschutz in allen 18 Vertragsstaaten der Europäischen Patentorganisation (EPO) erhalten. Die Europäische Patentorganisation ist eine zwischenstaatliche Organisation, die auf der Basis des 1977 in Kraft getretenen Europäischen Patentübereinkommens (EPÜ) gegründet wurde.

Amtssprachen

Die Amtssprachen des Europäischen Patentamts sind Deutsch, Englisch und Französisch. Europäische Patentanmeldungen sind in einer der Amtssprachen einzureichen.

Wege der Patenterteilung

Das europäische Patenterteilungsverfahren ist ein Prüfungsverfahren, dem eine Formalprüfung und eine obligatorische Recherche (der **erste Verfahrensabschnitt**) vorausgehen. Die Recherche dient der Bestimmung des aktuellen Stands der Technik, der Grundlage für die Beurteilung der Neuheit und der erfinderischen Tätigkeit einer Patentanmeldung ist.

Im Rahmen des Patenterteilungsverfahrens führen Ingenieure, die auf dem betreffenden Gebiet erfahren sind, einen solchen Vergleich mit dem in der Patent- und der technischen Literatur veröffentlichten Stand der Technik durch. Patentämter unterhalten deshalb eine umfangreiche technische Dokumentation, die eine umfassende Recherche ermöglicht. Die Sammlung des Europäischen Patentamts (EPA) umfaßt bereits mehr als 30 Millionen Schriftstücke und wächst jährlich um mehr als 500000 Dokumente.

Der **zweite Verfahrensabschnitt** umfaßt die Sachprüfung und gegebenenfalls die Patenterteilung. Eine Prüfungsabteilung setzt sich in der Regel aus drei technisch vorgebildeten Prüfern zusammen. Auf der Grundlage des in den Recherchenberichten festgelegten Stands der Technik müssen die Prüfer entscheiden, ob eine Patentanmeldung patentfähig ist, d. h. ob sie sich auf ein patentierbares Objekt bezieht, das neu, erfinderisch und gewerblich anwendbar ist. Am Ende des zweiten Verfahrensabschnitts kann die Patenterteilung stehen.

Ein angemeldetes oder ein erteiltes Patent kann angefochten werden. Das Einspruchsverfahren, an dem Dritte, d. h. Wettbewerber beteiligt sind, bildet den **dritten Verfahrensabschnitt**. Es wird von einer Einspruchsabteilung geprüft.

Einen **besonderen Verfahrensabschnitt** stellt das Beschwerdeverfahren dar. Während der drei vorgenannten Verfahrensabschnitte kann Beschwerde eingelegt werden gegen Entscheidungen, die von der Eingangsstelle, der Rechtsabteilung, den Prüfungsabteilungen oder den Einspruchsabteilungen in erster Instanz getroffen wurden. Die Beschwerdekammern entscheiden über die Beschwerden in zweiter und letzter Instanz. Die Mitglieder der Kammern sind unabhängig. Sie sind in ihren Entscheidungen an Weisungen nicht gebunden und nur dem Europäischen Patentübereinkommen (EPÜ) unterworfen.

Die Technischen Beschwerdekammern sind für Beschwerden gegen Entscheidungen über Zurückweisung europäischer Patentanmeldungen, gegen Entscheidungen über Erteilung von europäischen Patenten und gegen Entscheidungen der Einspruchsabteilungen zuständig. Die Technischen Beschwerdekammern setzen sich in der Regel aus zwei technisch vorgebildeten und einem rechtskundigen Mitglied zusammen.

In anderen Fällen, in denen nicht vorgesehen ist, daß technisch vorgebildete und rechtskundige Mitglieder gleichzeitg tätig werden, ist die Juristische Beschwerdekammer zuständig. Sie setzt sich nur aus rechtskundigen Mitgliedern zusammen.

Die Große Beschwerdekammer wird befaßt, wenn eine einheitliche Rechtsprechung sichergestellt werden soll oder wenn sich eine Rechtsfrage von grundsätzlicher Bedeutung stellt.

3.2 Das EPÜ-Korpus

Das EPÜ-Korpus ist ein Ausschnitt aus dem Text des Europäischen Patentübereinkommens (EPÜ). Das EPÜ verfolgt den Zweck, durch die Schaffung eines einheitlichen europäischen Patenterteilungsverfahrens auf der Grundlage eines einheitlichen materiellen Patentrechts den Schutz von Erfindungen in den Vertragsstaaten zu erleichtern, zu verbilligen und zu verstärken. Das EPÜ trat 1977 in Kraft und ist ein Sonderabkommen im Sinne des Pariser Verbandsübereinkommens (PVÜ) zum Schutz des gewerblichen Eigentums.

Das EPÜ umfaßt ca. 400 Seiten in 178 Artikeln. Es wurde parallel in den drei Amtssprachen verfaßt. Der dt., der engl. und der frz. Text sind jeweils rechtsgültig. Jeder Artikel hat eine Überschrift und ist in Absätze unterteilt, die meist numeriert sind.

Verschiedene Ausschnitte des EPÜ wurden für einen Test von Satzalignierungsverfahren verwendet (siehe Abschnitt 4.4 Evaluierung, Seite 49).

3.3 Das EBK-Korpus

Das EBK-Korpus ist das Korpus der Entscheidungen der Beschwerdekammern des Europäischen Patentamts. Die Beschwerdekammern entscheiden in zweiter und letzter Instanz über die dort vorgetragenen Beschwerden. Die Entscheidungen werden in schriftlicher Form in einer der drei Amtssprachen (Deutsch, Englisch oder Französisch) von den Mitgliedern einer Beschwerdeabteilung niedergelegt und anschließend meist in die zwei weiteren Amtssprachen übersetzt (vom Sprachendienst des Europäischen Patentamts).

Das EBK-Korpus umfaßt ca. 5 000 Entscheidungen (12 Millionen Wörter) aus dem Zeitraum 1979 bis 1990. Die Originalsprache der Dokumente ist Englisch (53 %), Deutsch (37 %) oder Französisch (10 %). Da die Entscheidungen vom Sprachendienst nach Dringlichkeit übersetzt werden, liegen weniger als ein Drittel der Entscheidun-

gen in allen drei Sprachen vor, ein größerer Teil existiert in einer engl. und dt. Version. Das engl. Korpus hat eine Größe von ca. 5,5 Mio. Wörtern, das dt. eine Größe von ca. 4,5 Mio Wörtern und das frz. eine Größe von ca. 2 Mio. Wörtern.

3.3.1 Strukturierung eines Dokuments

Jedes Dokument ist in Einheiten unterteilt, die mit Markierungen (*Tags*) annotiert sind. Eine Markierung steht am Anfang der Zeile. Sie wird durch die Zeichen "<" bzw. ">" begrenzt und besteht immer aus drei Buchstaben. Eine Einheit kann eine oder mehrere Zeilen (abgetrennt durch Zeilenvorschub) enthalten. Eine Zeile kann aus formalen Angaben (siehe unten) oder aus einem oder mehreren Sätzen bestehen.

Ein Dokument läßt sich automatisch in vier Teile zerlegen, deren Längen innerhalb eines Dokuments und von einem Dokument zum andern sehr stark variieren.

Die vier Teile eines Dokuments

- 1. Teil: Formale Angaben zur Entscheidung

  Dazu gehört u. a. die Dokumentnummer (die Dokumente sind fortlaufend numeriert), die Nummer des Falls (<CSN> von *case number*), der Name des Einsprechenden (<OPP> von *opponent*), die Sprachkennung, die Originalsprache, der Kode der Internationalen Patentklassifikation IPC, das Datum, der oder die Artikel des EPÜ, auf die sich die Entscheidung bezieht, Schlagwörter (<KEY> von *keyword*), Stichwörter (<HDW> von *headword*) und Leitsätze (<HDN> von *headnote*).

- 2. Teil: Sachverhalt und Anträge

  Hier sind alle Einheiten mit der Markierung <FSU> (von: *facts and submissions*) versehen. Dieser Teil ist durch Nummern weiter strukturiert; es wurden jedoch Abweichungen dieser Numerierung zwischen den einzelsprachlichen Versionen eines Dokuments beobachtet (dies gilt auch für den dritten Teil).

- 3. Teil: Entscheidungsgründe

  Hier sind alle Einheiten mit der Markierung <RES> (von: *reasons for the decision*) versehen.

- 4. Teil: Entscheidungsformel

  Hier sind alle Einheiten mit der Markierung <ORD> (von: *order*) versehen.

3.3.2 Erstellung von Teilkorpora

Im Verlauf der Studie wurden verschiedene Teilkorpora zu unterschiedlichen Zwecken zusammengestellt. Für die Evaluierung der Satzalignierung waren Kriterien wie

Dokumentlänge, durchschnittliche Länge eines Abschnitts, durchschnittliche Satzlänge usw. relevant.

Für eine terminologische Untersuchung ist es wichtig, Dokumente ähnlichen Inhalts in Teilkorpora zusammenzustellen, um zu gewährleisten, daß sich ein Teil des Wortschatzes eines Dokuments mit dem Wortschatz eines weiteren Dokuments (oder mehrerer) überschneidet. Es stellte sich heraus, daß diese Aufgabe schwieriger war, als ein erster Blick auf die Dokumente es erwarten ließ. Die Mitglieder der Beschwerdekammer ordnen einem Dokument Schlagwörter und Stichwörter zu, die sich auf juristische und/oder technische Inhalte des Dokuments beziehen. Tendenziell beziehen sich die Stichwörter eher auf technische Sachverhalte (z. B. Katalysator, Druckmaschine, Theta-1), und die Schlagwörter eher auf juristische Inhalte (z. B. Neuheit von Sachansprüchen, Ergänzung der spezifischen Offenbarung durch allgemeines Wissen, Automatischer Widerruf usw.). Eine genauere Prüfung der Stichwörter und der Schlagwörter von 300 Dokumenten zeigte aber, daß sich diese Tendenz nicht verallgemeinern läßt. Deshalb wurden nur Entscheidungen der Technischen Beschwerdekammer 3.3.1 (Chemie) herangezogen, aus denen aufgrund korpuslinguistischer Kriterien ein Teilkorpus zusammengestellt wurde. Dieses Teilkorpus enthält sowohl eine ausreichende Anzahl dreisprachiger Dokumente (151) als auch eine ausreichende Anzahl längerer Dokumente (wichtig für die Tests zur Satzalignierung).

3.4 Eigenschaften der Texte aus der Patentdokumentation

Die Entscheidungen der Beschwerdekammern in **drei Sprachen** (EBK-Korpus) sind nur eine von mehreren Textsorten aus der Patentdokumentation. Sie sind für eine computerlinguistische Untersuchung unter dem Aspekt der Übersetzung besonders gut geeignet, da sie

- eine stark differenzierte und einheitliche Strukturierung aufweisen,
- sehr umfangreich sind (ein Erfordernis für die Anwendung statistischer Methoden),
- in einem juristischen Kontext verfaßt wurden, so daß die Übersetzungen weitgehendst standardisiert sind.

An dieser Stelle möchte ich zwei Arbeiten erwähnen, in denen andere Textsorten aus der Patentdokumentation (dt. Patent- und Auslegeschriften) unter lexikalisch-syntaktischen bzw. textlinguistischen Aspekten untersucht wurden. Über die informationslinguistische Untersuchung von Patentschriften (PADOK-Projekt) wird an anderer Stelle (siehe Abschnitt 6.2 Das PADOK-Projekt, Seite 113) berichtet.

Wortbildung in Patenttexten

Die Arbeit von Dederding[1] beschäftigt sich mit dem Zusammenhang von Text und Wortbildung in Patenttexten. Dederding nimmt an, daß in solchen Texten besonders

1. Dederding 1982.

viele Neubildungen[1] von Komposita vorkommen und sich diese Neubildungen aus dem Kontext entwickeln, da es sich bei Patentschriften um Beschreibungen von Erfindungen handelt, die technische Neuerungen darstellen. Diese Ausgangshypothese erwies sich jedoch nicht als zutreffend: Die Analyse von Neubildungen ist kein spezifisches Problem der Patenttexte.

Textlinguistische Besonderheiten in Patentschriften

Schamlu[2] hat eine argumentationstheoretische Analyse der Textsorte Patentschrift und damit auch des Patenterteilungsprozesses durchgeführt und geht dabei auch auf den Zusammenhang zwischen syntaktischen Strukturen und Argumentationszielen in den Patentschriften ein. In diesem Zusammenhang ist es interessant, daß sie auf die **komplexe syntaktische Struktur der Nominalphrasen** hinweist[3], ein Sachverhalt, der auch in den Texten des EBK- und des EPÜ-Korpus zu beobachten war und der zu erheblichen Problemen bei der maschinellen Aufbereitung der Texte führte.

1. Der Ansatz der Arbeit liegt dort, wo lange Zeit der Schwerpunkt der Wortbildungsforschung lag: im Verhältnis von Wortbildungen und den ihnen “entsprechenden” syntaktischen Strukturen. Die Untersuchung beschränkt sich auf Nominalkomposita, worunter Wortbildungskonstruktionen verstanden werden, die mindestens zwei Wortstämme enthalten und als ganzes der Wortart Substantiv angehören. Wortbildungen dieses Typs und ihre Bestandteile sind maximal markant, d. h. wenn es um das Aufsuchen von Entsprechungen zu Wortbildungselementen geht, sind die den Nominalkomposita entsprechenden Elemente leichter auffindbar als etwa Entsprechungen zu Suffixen oder Präfixen.

2. Schamlu 1985.

3. Schamlu nennt die komplexe syntaktische Struktur des Patentanspruchs neben der hohen Abstraktionsebene als zweiten Hauptgrund für die Unverständlichkeit der Patentschriften: “Linguistisch gesehen, besteht der Patentanspruch ... aus einer komplexen Nominalphrase mit zahlreichen modifizierenden Relativsätzen, die teilweise ineinander verschachtelt sind. Dies erschwert die Dekodierung in dem Maße, daß nur eine detaillierte Aufgliederung zum Verständnis führt. Die Dekodierung wird außerdem durch häufig unklare syntaktische Bezüge zwischen den Konstituenten behindert” (Schamlu 1985, S. 177).

4 Satzalignierung

Im folgenden Kapitel soll zunächst erklärt werden, was Alignierung ist und auf welchen Ebenen sie angewandt werden kann. Nach einem allgemeinen methodischen Überblick werden zwei Ansätze zur Satzalignierung näher beschrieben, für die eine Implementierung vorlag[1]. Beide Methoden wurden unter verschiedenen Aspekten getestet und evaluiert. Abschließend werden einige Anwendungen der Satzalignierung vorgestellt.

4.1 Grundlagen

4.1.1 Aufgabe der Alignierung

Ein Alignierungsprozeß nimmt sog. **parallele Texte** als Eingabe, d. h. einen Quelltext und die Übersetzung dieses Textes[2] in eine oder mehrere Sprachen.

Einen Text mit seiner Übersetzung zu alignieren heißt, zu zeigen, welche Teile dieses Textes durch welche Teile des zweiten Textes übersetzt sind. Alignierung gibt es auf verschiedenen Ebenen - Kapitel, Abschnitte, Absätze, Sätze, Phrasen, Wörter - mit unterschiedlich ausgeprägter Sequentialität. So ist z. B. die Sequentialität auf der Ebene der Abschnitte viel höher als die Sequentialität auf der Ebene der Phrasen oder der Wörter.

Das Ergebnis des Alignierungsprozesses wird dargestellt in Form einer Liste von Einheiten aus den beiden Texten - z. B. Sätzen - und ihren Entsprechungen. Die Aufgabe der Satzalignierung besteht darin, Entsprechungen zwischen den Sätzen im Quelltext und den Sätzen im Zieltext zu finden. Eine Lösung des Alignierungsproblems besteht in der Untermenge des karthesischen Produkts der Menge der Sätze im Quelltext und der Menge der Sätze im Zieltext.[3] Die Ausgabe zeigt die Alignierungen zwischen den Sätzen. In den meisten Fällen entspricht ein Satz des Quelltextes genau einem Satz des Zieltextes. Es gibt jedoch auch kompliziertere Entsprechungen, die später noch genauer erörtert werden.

Die Satzalignierung ist ein erster Schritt bei dem überaus schwierigen Unterfangen, Entsprechungen zwischen Wörtern oder Phrasen zu finden. Im folgenden wird *Alignierung* als eine Kurzform für *Satzalignierung* verwendet.

1. Wallner 1994, Blank 1995.

2. Im folgenden wird bei parallelen Texten immer ein Text als *Quelltext* und ein Text als *Zieltext* bezeichnet, unabhängig davon, welcher Text im einzelnen Fall der Quelltext oder der Zieltext ist.

3. Diese Annahme gilt für den allgemeinsten Fall: ein Satz wird durch genau einen Satz übersetzt.

4.1.2 Beispiel eines parallelen Texts

Ausschnitt aus Dokument 103 (dt. Text) und Dokument 104 (engl. Text) des EBK-Korpus. (Vor jedem Satz steht eine Satznummer).

Tabelle 1. Beispiel eines parallelen Texts

engl. Text	**dt. Text**
8: <FSU> Both parties filed initial observations dated 1 and 23 February 1989, respectively, and further observations in reply dated 30 May and 5 June 1989.	8: <FSU> Beide Beteiligte reichten am 1. bzw. 23. Februar 1989 eine erste und am 30. Mai bzw. am 5. Juni 1989 eine weitere Stellungnahme ein.
9: Both parties requested oral proceedings under Article 116 EPC.	9: Beide Parteien beantragten eine mündliche Verhandlung nach Artikel 116 EPÜ.
10: <FSU> III. As to question (i): In the communication dated 14 October 1988, it was suggested that the "protection conferred" by a patent is to be determined in accordance with Article 69 EPC and its Protocol, and is distinct from the "rights conferred" by a patent which are to be determined by individual national laws of the designated Contracting States in accordance with Article 64(1) EPC.	10: <FSU> III. Zu Frage i: 11: <FSU> In dem Bescheid vom 14. Oktober 1988 wurde ausgeführt, daß der "Schutzbereich" eines Patents entsprechend Artikel 69 EPÜ und dem dazugehörigen Protokoll festzulegen sei und daß er sich von den "Rechten aus dem Patent" unterscheide, die sich gemäß Artikel 64 (1) EPÜ aus dem nationalen Recht der benannten Vertragsstaaten ergäben.
11: Accordingly, under Article 123(3) EPC the question to be considered is whether the matter which is protected by the claim, as defined by its technical features, is extended.	12: Dementsprechend müsse im Hinblick auf Artikel 123 (3) EPÜ geprüft werden, ob der durch den Anspruch geschützte und durch seine technischen Merkmale definierte Gegenstand erweitert werde.
12: The Appellant submitted that rigid lines of demarcation between categories of claims and the protection thereby conferred did not exist, and that the considerations under Article 123(3) EPC were as set out by the Enlarged Board in the communication.	13: <FSU> Die Beschwerdeführerin behauptete, es gebe keine starren Grenzen zwischen den verschiedenen Anspruchskategorien und dem durch sie gewährten Schutz. 14: Ihre Überlegungen zu Artikel 123 (3) EPÜ entsprächen denen im Bescheid der Großen Beschwerdekammer.

Tabelle 1. Beispiel eines parallelen Texts

engl. Text	dt. Text
13: The Respondent submitted that a "use" claim in respect of an article is narrower in scope than an original "article" claim; and that it was not necessary to consider national laws concerning infringement, for the reason set out in the above communication.	15: <FSU> Die Beschwerdegegnerin trug vor, daß ein "Verwendungsanspruch" für ein Erzeugnis vom Schutzumfang her enger sei als ein reiner "Erzeugnisanspruch"; es sei aus dem in dem Bescheid genannten Grund nicht erforderlich, das nationale Verletzungssrecht zu berücksichtigen.
14: <FSU> IV. As to question (ii): The Appellant relied upon Decision T 378/86 (OFJ EPO, 1988, 386) in support of his submission that the extent of protection conferred by a "product" claim encompasses that conferred by a "use" claim, and further submitted that amendment from a per se product claim to a use claim was therefore a disclaimer.	16: <FSU> IV. Zu Frage ii: 17: <FSU> Die Beschwerdeführerin berief sich auf die Entscheidung T 378/86 (ABl. EPA 1988, 386), um ihre Behauptung zu stützen, daß der Schutzbereich eines "Erzeugnisanspruchs" den eines "Verwendungsanspruchs" einschließe und daß eine Änderung eines reinen Erzeugnisanspruchs in einen Verwendungsanspruch deshalb einen Disclaimer darstelle.
15: The Respondent agreed that such an amendment did not contravene Article 123(3) EPC, and was allowable provided that such use is disclosed in the patent specification as originally filed and as granted, and that the use is both novel and inventive.	18: <FSU> Die Beschwerdegegnerin schloß sich der Auffassung an, daß eine solche Änderung nicht gegen Artikel 123 (3) EPÜ verstoße und somit zulässig sei, sofern die Verwendung in der ursprünglich eingereichten und erteilten Fassung der Beschreibung offenbart und sowohl neu als auch erfinderisch sei.
16: <FSU> V. As to question (iii): (a) In his initial observations, the Appellant submitted that the intended use or purpose as expressed in the claims confers novelty on them, in accordance with a line of authority developed by decisions of the Boards of Appeal, in accordance with the EPC.	19: <FSU> V. Zu Frage iii: 20: <FSU> a) In ihrer ersten Stellungnahme führte die Beschwerdeführerin aus, daß die in den Ansprüchen ausgedrückte beabsichtigte Verwendung oder der beabsichtigte Zweck den Ansprüchen nach dem EPÜ entsprechend der Rechtsprechung der Beschwerdekammern Neuheit verleihe.
17: The submissions on behalf of the Appellant in case G 6/88 were adopted by the Appellant.	21: <FSU> Die Beschwerdeführerin schloß sich dem Vorbringen der Beschwerdeführerin im Fall G 6/88 an.

4.1.3 Verfahren zur Satzalignierung

Verschiedene Forschungsgruppen beschäftigen sich mit Satzalignierung und erzielen dabei sehr gute Ergebnisse. Die Alignierung von Sätzen vollzieht sich immer in einer Abfolge von zwei Schritten: Zuerst werden Ähnlichkeitswerte für die verschiedenen möglichen Entsprechungen berechnet, danach werden die Kombinationen bestimmt, die die Ähnlichkeit für den gesamten Text maximieren.

Je nach Ähnlichkeitswert, der für die Sätze berechnet wird, lassen sich, grob genommen, zwei Methoden unterscheiden: eine **satzlängenbasierte** Methode einerseits und eine **linguistische/lexikalische** Methode andererseits. Es gibt auch Verfahren (z. B. Gaussier 1995), die auf einer Kombination der beiden Ansätze basieren.

Eine der beiden Methoden - die satzlängenbasierte - arbeitet mit den formalen Eigenschaften der zu analysierenden Texte; dabei wird die Satzlänge entweder aus der Anzahl der Wörter[1] oder aus der Anzahl der Buchstaben[2] bestimmt.

Die zweite Methode - die linguistische bzw. lexikalische - basiert auf der Alignierung der lexikalischen Einheiten, aus denen der Satz besteht und ist deshalb mehr inhaltsbezogen. Sie wird von Kay und Röscheisen (1988), Catizone et al. (1989) und anderen Gruppen angewandt. Debili und Sammouda (1992) konzipierten einen Alignierungsalgorithmus, der auf Entsprechungen zwischen Wörtern basiert; diese Entsprechungen werden mit Hilfe eines zweisprachigen Wörterbuchs festgelegt. Ihr Vorgehen ähnelt dem von Catizone et al. (1989). Jede der beiden Methoden wird anhand eines Algorithmus genauer vorgestellt.

4.2 Die Methode von Kay und Röscheisen

Der Alignierungsalgorithmus von Kay und Röscheisen basiert nur auf interner Evidenz, d. h. auf der Feststellung, daß ein Paar von Sätzen, das ein aligniertes Paar von Wörtern enthält, selbst aligniert sein muß. Das Grundkonzept des Algorithmus ergibt sich aus den folgenden Beobachtungen:

- Die Alignierung von Sätzen basiert auf der Alignierung von Wörtern.
- Der Algorithmus ist ein Relaxationsprozeß, bei dem eine partielle Alignierung auf der Ebene der Wörter dazu benutzt wird, die wahrscheinlichste Alignierung auf der Ebene der Sätze zu bestimmen. Diese Alignierung auf Satzebene wird in der nächsten Iteration zu einer Verfeinerung der Alignierung auf Wortebene benutzt.

Der Algorithmus beginnt mit einer initialen Untermenge von alignierbaren Sätzen unter Ausschluß der Sätze, deren relative Positionen im Text so unterschiedlich sind, daß die Wahrscheinlichkeit, daß diese Sätze einander entsprechen, extrem niedrig ist.

1. Brown et al. 1990.
2. Gale & Church 1991.

Die initiale Menge potentiell alignierbarer Sätze wird mit **Ankerpunkten** berechnet, das sind Satzpaare, die in jedem Fall aligniert sind (Anfang und Ende der Texte, Abschnittsgrenzen usw.).

Dann wird eine Menge von möglichen Wortalignierungen berechnet, indem Wörter ausgewählt werden, die in Sätzen auftreten, die möglicherweise miteinander aligniert sind. Die Wörter eines Wortpaares gelten dann als Kandidaten für die Wortalignierung, wenn sie ähnliche Verteilungen im jeweiligen Text haben. Die Verteilung der Wörter eines Wortpaars ist dann ähnlich, wenn die meisten Sätze, in denen das erste Wort auftritt, alignierbar sind mit Sätzen, in denen das zweite Wort auftritt. Die offensichtlich zuverlässigsten Wortalignierungen werden dann dazu verwendet, eine neue Abschätzung der Satzalignierung vorzunehmen. Da Satzalignierungen nie aus der Menge der Alignierungen entfernt werden, konvergiert der Prozeß in einem Punkt, an dem keine neuen Alignierungen mehr gefunden werden können, und der Algorithmus terminiert.

Die zur Verarbeitung notwendigen Informationen werden in Tabellen gehalten. Die Tabellen werden in einem Schritt generiert und im nächsten Schritt als Ausgangsdaten verwendet. Der Algorithmus arbeitet mit vier Hilfstabellen:

Wort-Satz-Index (*word sentence index*: WSI)

Der WSI gibt für jedes Wort eines Textes an

- in welchen Sätzen das Wort enthalten ist (dargestellt durch Satznummern)
- wie oft das Wort im gesamten Text auftritt (absolute Häufigkeit).

Tabelle der alignierbaren Sätze (*alignable sentence table*: AST)

Hier werden alle Kombinationen von Sätzen eingetragen, die aufgrund ihrer jeweiligen Position im Text möglicherweise Übersetzungen voneinander sind. Die Tabelle wird in jedem Relaxationsschritt kleiner, da die Ergebnisse des vorherigen Relaxationsschritts sich auf die nachfolgenden Kombinationsmöglichkeiten beschränkend auswirken.

Wort-Alignierungs-Tabelle (*word alignment table*: WAT)

Die WAT ist eine Liste derjenigen Wörter, die sich möglicherweise entsprechen, da sie ähnliche Verteilungen in den beiden Texten haben. Der Grad der Ähnlichkeit wird mit einer Formel (siehe unter Abschnitt 4.2.1 Der Relaxationsprozeß, Seite 36) berechnet. Die Liste enthält die Wortpaare zusammen mit Angaben über die Ähnlichkeit der Wörter untereinander und ihre Frequenz. Die WAT wird in jedem Relaxationsschritt neu generiert.

Satz-Alignierungs-Tabelle (*sentence alignment table*: SAT)

Die in der WAT dokumentierten Wortalignierungen lassen Rückschlüsse auf mögliche Satzalignierungen zu. Diese möglichen Satzalignierungen werden in der SAT abgelegt. Dabei wird auch eingetragen, wie viele Wortalignierungen für das jeweilige Satzpaar sprechen.

4.2.1 Der Relaxationsprozeß

Jeder Durchgang des Relaxationsprozesses erzeugt auf der Grundlage des WSI und einer jeweils aktualisierten AST eine neue WAT und SAT.

Satzalignierungen aus der SAT, die sich auf eine Mindestanzahl von Wortalignierungen stützen, werden vom Programm zu **neuen Ankerpunkten** gemacht. Sie sind die Grundlage für die Generierung der AST im nächsten Schritt. Ein Satzpaar ist dann ein Ankerpunkt, wenn die Sätze, die es konstituieren, mit hoher Wahrscheinlichkeit Übersetzungen voneinander sind. Solche Satzpaare stellen feste Verbindungen zwischen den beiden Texten her. Das hat zur Folge, daß in den nachfolgenden Generierungen der AST keine Satzpaare erzeugt werden dürfen, die diese bereits existierenden Ankersatzpaaare überkreuzen würden. Eine Satzzuordnung, die einmal als Anker bestimmt wurde, wird im weiteren Verlauf des Prozesses nicht mehr aufgelöst. Das Relaxationsverfahren terminiert, wenn aus der SAT keine weiteren Anker mehr hervorgehen.

4.2.1.1 Bestimmung der alignierbaren Sätze

Die AST enthält die für eine Alignierung in Frage kommenden Satzpaare. Als Modell kann man sich das karthesische Produkt aus den Mengen der Sätze des Quell- und des Zieltexts vorstellen.

In parallelen Texten müssen sich die korrespondierenden Sätze innerhalb eines bestimmten Bereichs befinden, so daß es nicht notwendig ist, jeden Satz des Quelltexts mit jedem Satz des Zieltexts zu kombinieren. Beispielsweise sind die beiden Anfangssätze und die beiden Schlußsätze mit hoher Wahrscheinlichkeit jeweils Übersetzungen voneinander. Im mittleren Bereich der Texte ist es schwieriger, die richtigen Zuordnungen vorzunehmen, zumal, wenn man berücksichtigt, daß es nicht nur 1:1-Entsprechungen zwischen den Sätzen gibt, sondern auch 1:0-, 0:1-, 2:1-Entsprechungen usw. möglich sind. Je weiter die Sätze von den Ankern entfernt liegen, desto weniger können die Positionen der korrespondierenden Sätze bestimmt werden. Daher müssen für die Satzkombinationen in der Mitte zwischen zwei benachbarten Ankern breitere Bereiche angelegt werden. Kay und Röscheisen berücksichtigen diese Bedingungen bei der Berechnung der AST folgendermaßen:

s_i (i=1 ... m) seien die Sätze des Quelltexts

s_j (j=1 ... n) seien die Anzahl der Sätze im Zieltext,

so kann man sich die Erstellung der AST als *m*×*n*-Matrix vorstellen, deren Felder entsprechend gekennzeichnet sind, je nachdem, ob an an dieser Stelle eine Satzalignierung möglich oder unmöglich ist. Jedem Satz im Quelltext entsprechen in der AST durchschnittlich n/m Sätze des Zieltexts.

Die Erwartungswerte für die Sätze liegen somit auf der Geraden, die durch die nachfolgende Gleichung beschrieben wird:

$$j = i\,(\frac{n}{m})$$

Die maximale Abweichung σ befindet sich an der Stelle j=n/2, d. h. in der Mitte zwischen den Ankern und soll $\sigma = \sqrt{n}$ betragen. So läßt sich zu jedem Satz des Ausgangstexts ein Intervall ([*mittlerer Erwartungswert - Abweichung, mittlerer Erwartungswert + Abweichung*]) berechnen, in dem die Sätze liegen, die möglicherweise miteinander korrespondieren.

Für die Berechnung der jeweiligen Abweichung ergibt sich die folgende Formel:

$$\sigma = \begin{array}{c} \sqrt{j},\, fuer\,(1 \le i \le \frac{m}{2}) \\ \sqrt{n-j},\, fuer\,(\frac{m}{2} \le i \le m) \end{array}$$

Der Bereich, der einem Satz s_i zuzuordnen ist, ergibt sich somit aus dem Intervall [j - σ; j + σ].

Die **initiale AST** wird vor dem ersten Durchlauf des Relaxationsprozesses erstellt. Ihre Generierung unterscheidet sich von den weiteren dadurch, daß nur zwei Anker zur Verfügung stehen. Das sind:

1. Satzpaar 1, gebildet aus dem jeweils ersten Satz des Quell- und des Zieltexts.
2. Satzpaar 2, gebildet aus dem jeweils letzten Satz des Quell- und des Zieltexts.

Bei der Generierung der späteren AST’s wird die beschriebene Methode jeweils auf Textbereiche zwischen je zwei Ankern angewendet.

4.2.1.2 Berechnung der Wortähnlichkeit

Paare von Wörtern, die in die WAT eingetragen werden, müssen eine gewisse Ähnlichkeit aufweisen.

Für alle Satzkombinationen (s^A, s^B)[1] der AST wird jedes Wort $v \in s^A$ mit jedem Wort $w \in s^B$ kombiniert, um aufgrund eines Ähnlichkeitsmaßes zu bestimmen, ob die Wörter des Wortpaares (*v, w*) möglicherweise eine Übersetzung voneinander sind. Das Ähnlichkeitsmaß SIM (*similarity*) ist:

$$SIM = \frac{2c}{N_A(v) + N_B(w)}$$

N_A(v): Häufigkeit von Wort v in $Text_A$

N_B(w): Häufigkeit von Wort w in $Text_B$

c: Anzahl von gemeinsamen Auftreten des Wortpaares (v, w) in Satzpaaren der AST

Eintrag der Wortpaare und Wortähnlichkeiten in die WAT

Die WAT kann mit dem Kriterium der Häufigkeit in mehrere Segmente, i. a. zwei bis drei, unterteilt werden. Dadurch erhält man unterschiedlich zuverlässige WAT-Segmente, die eine prioritätsgesteuerte Verarbeitung der Wortalignierungen ermöglichen. Die Wortalignierungen innerhalb eines Segments werden nach ihren Ähnlichkeitswerten sortiert. Nur Wortpaare, deren Ähnlichkeit über einem vordefinierten Schwellenwert liegen, werden in die WAT eingetragen.

1. s^A sei ein Satz aus dem Quelltext, s^B ein Satz aus dem Zieltext.

Beispiel einer engl.-dt. Wortalignierungstabelle

Tabelle 2. Wort-Alignierungstabelle (Auszug aus der WAT von Dok. 103/104)

engl. Wort	dt. Wort	$N_A(v)$	$N_B(w)$	SIM
Article	Artikel	70	51	77
purpose	Zweck	42	29	67
question	Frage	38	27	58
function	Funktion	4	5	88
description	Beschreibung	9	8	82
defined	definiert	8	8	75
Board	Beschwerdekammer	36	31	89
Board	Großen	36	19	65
Enlarged	Beschwerdekammer	26	31	87
Enlarged	Großen	26	19	80
feature	Merkmal	30	27	94
feature	technisch	30	14	63

Erstellung der SAT

Nach der Eintragung der Wortähnlichkeiten in die WAT wird die SAT generiert. Die Satzpaare, die die zuvor zugeordneten Wortpaare enthalten, werden in die SAT eingetragen. Mit jedem Wortpaar, das für ein bestimmtes Satzpaar “spricht”, wird der Zähler dieses Satzpaares erhöht. Nachdem die SAT vollständig generiert ist, wird aufgrund dieses Zählers bestimmt, ob ein Satzpaar zu einem neuen Anker wird.

Bei der Generierung der SAT werden die Wortpaare nach absteigender Ähnlichkeit abgearbeitet. Für jedes Wortpaar (*v, w*) erfolgt ein Eintrag des dazugehörigen Satzpaares, wenn folgende Bedingungen erfüllt sind:

1. Das Satzpaar darf nur dann eingetragen werden, wenn es in der AST enthalten ist.

2. Das Satzpaar darf nur dann eingetragen werden, wenn es sich nicht mit bereits in der SAT existierenden Satzpaaren überkreuzt.

4.2.2 Verbindung zum EM-Algorithmus

Der beschriebene Relaxationsprozeß kann als Instantiierung des EM-Algorithmus[1] betrachtet werden, eines allgemeinen Schemas zur *Maximum-Likelihood*-Berechnung

1. Dempster et al. 1977.

aus unvollständigen Daten. Dieser Algorithmus ist eine Iteration von zwei Schritten, der Estimation (*estimation: E-Step*) und der Maximierung (*maximization*: *M-Step*).

- **Estimation**: Eine Menge von beobachtbaren Fakten dient als Basis für die Abschätzung einer anderen, nicht-beobachtbaren Menge. Im Fall der Satzalignierung heißt das: Unter Vorgabe einer AST und der Wortalignierungen in der WAT, soll für jeden Satz *i* der entsprechende Satz *j* im Zieltext nach dem oben beschriebenen Verfahren abgeschätzt werden.
- **Maximierung**: Die bisherigen Abschätzungen sind die Grundlage für eine neue Abschätzung, die zwar differenzierter ist, aber die alte Abschätzung bestätigt. Im Fall der Satzalignierung heißt das: es wird eine neue WAT erstellt, die die Wahrscheinlichkeiten der zuvor berechneten Satzalignierungen maximiert.

4.2.3 Die Implementierung

Der Originalalgorithmus wurde mit einigen Modifikationen und Verbesserungen von Wallner (1994) in Prolog II implementiert.

4.2.3.1 Programmstruktur

Das Programm besteht aus zwei Modulen: Vorverarbeitung und Alignierung.

Das Modul für die Vorverarbeitung generiert die WSI's und berechnet Frequenzinformationen für bestimmte Wörter (zusätzliche Ankerpunkte, Funktionswörter, Abschnittsgrenzen). Ein Teil dieser Berechnungen sind echte Neuerungen im Vergleich zum Originalalgorithmus[1].

Das Alignierungsmodul liest zuerst Parameter ein, die vom Benutzer zu definieren sind. Diese Parameter legen Schwellenwerte für die minimale Wortähnlichkeit und die minimale Zahl der Vorkommen eines Wortes fest. Sie können das Alignierungsergebnis beträchtlich verändern. Das Alignierungsmodul steuert den gesamten Relaxationsprozeß und ruft die entsprechenden Funktionen für die Generierung der beschriebenen Tabellen auf. Nach jedem Durchgang wird geprüft, ob ein neuer Durchgang nötig ist oder ob das Programmende erreicht ist.

4.2.3.2 Erweiterungen

Der von Kay und Röscheisen (1988) beschriebene Algorithmus wurde zur Verbesserung seiner Leistung unter verschiedenen Aspekten modifiziert. Dazu einige Beispiele:

- Vordefinierte **Abschnittsmarkierungen** (Anfang bzw. Ende eines Abschnittes) werden berücksichtigt. Ist die Anzahl der Abschnitte in beiden Texten gleich, so dienen die Abschnittsmarkierungen als zusätzliche Ankerpunkte bei der Generie-

1. Vgl. dazu Wallner 1994, S. 60.

rung der AST. Wenn die Anzahl der Abschnitte in den beiden Texten nicht gleich ist, dann aligniert das Programm die Texte, als ob diese Abschnittsmarkierungen nicht existierten. In diesem Punkt ist der Kay-Röscheisen-Algorithmus leistungsstärker als der Church-Gale-Algorithmus, der in einem solchen Fall überhaupt keine Alignierung vornimmt.

- Vordefinierte sprachspezifische **Funktionswörter** werden erkannt. Für diese werden keine Wortähnlichkeitswerte berechnet, da sie zu häufig sind, als daß sie als Basis für die Satzalignierung dienen könnten. Außerdem wird der Algorithmus schneller, wenn er für diese Wörter keine Berechnungen ausführt.
- Das Programm untersucht den Text nach **orthographisch identischen oder ähnlichen Wörtern** (*cognates*). Wenn solche Wörter in beiden Texten mit derselben Frequenz und in miteinander alignierbaren Textsegmenten vorkommen, werden sie als zusätzliche Ankerpunkte verwendet.
- Für eine schnellere Programmausführung wurde die **Generierung der WAT** verändert, sie verwendet aber immer noch dieselbe Formel. Die **Generierung der AST** wurde verändert, so daß die Standardabweichung kleiner wird. Diese modifizierte Version der AST verbesserte nachweislich die Ergebnisse der Alignierung[1].

4.2.4 Evaluierung

Der Algorithmus wurde mit denselben wissenschaftlichen Texten (engl.-dt.) getestet, die Kay und Röscheisen in ihrer Evaluation verwendet hatten[2].

Ziel der Testreihen

Die Tests von Wallner[3] wurden so konzipiert, daß beurteilt werden konnte, wie sich verschiedene Parameter wie Textlänge, vordefinierte Abschnittsmarkierungen usw. gegenseitig beeinflussen. Sie gliedern sich in sechs Testreihen, die folgende Aufgaben erfüllen:

- Tests des Alignierungsprogramms mit unterschiedlichen Textlängen und bei gleicher Parametereinstellung
- Tests mit und ohne feststehende Ankerpunkte: Abschnittsmarkierungen, Eigennamen, Zahlen
- Tests mit breiter und schmaler AST
- Tests mit variierten Benutzerparametern und gleichbleibender Textlänge
- Test mit und ohne Berücksichtigung von Funktionswörtern

1. Vgl. dazu Wallner 1994, S. 60.
2. Engl. Text: Cosmic Rays from Cygnus X-3 (*Scientific American*, Vol. 253, Nummer 5, November 1985), dt. Text: Kosmische Strahlen von Cygnus X-3. (*Spektrum der Wissenschaft*, Januar 1986).
3. Vgl. dazu Wallner 1994, S. 82 ff.

- Test des Alignierungsprogramms unter Verwendung monolingualer Lexika.

Kriterien für die Beurteilung der Ergebnisse

Die vom Programm alignierten Texte wurden manuell korrigiert und nach folgenden Kriterien bewertet:

1. Korrektheit der SAT (*correctness in SAT*): Anzahl der richtig alignierten Satzpaare in der SAT im Verhältnis zur Anzahl der Gesamteinträge in der SAT.

2. Abdeckung der SAT (*coverage in SAT*): Anzahl der unterschiedlichen Satznummern, die mindestens einmal in der SAT auftreten, im Verhältnis zu der Gesamtanzahl aller Sätze beider Texte.

3. absolute Korrektheit: Anzahl der richtig erkannten Satzzuordnungen.

Ergebnisse

Die Ergebnisse variieren in Abhängigkeit von den erwähnten Kriterien. Nimmt man die höchste und die niedrigste erzielte absolute Korrektheit als Ausgangspunkt, erhält man folgendes Ergebnis[1]:

Tabelle 3. Ergebnisse der Alignierung mit der Kay-Röscheisen-Methode

	Korrektheit der SAT	**Abdeckung**	**absolute Korrektheit**
bester Fall	89,12 %	93,30 %	89,12 %
schlechtester Fall	100 %	18,77 %	16,67 %

Die Testreihen zeigten, daß

- die Textlänge die Korrektheit der Alignierung erheblich beeinflußt (die besten Ergebnisse wurden mit einer Textlänge von mehr als 250 Sätzen erzielt).
- zusätzliche Ankerpunkte die Korrektheit der Alignierung verbessern.
- der Ausschluß von Funktionswörtern den Algorithmus beschleunigt und die Korrektheit erhöht.
- es besser ist, eine AST mit einer kleineren Abweichung zu verwenden als die von Kay und Röscheisen verwendete AST.

1. Vgl. dazu Wallner 1994, Kap. 6.

- die vom Benutzer definierten Schwellenwerte für die Wortähnlichkeit und die minimale Zahl der Wortvorkommen[1] die Vollständigkeitsquote (*recall*[2]) und die Genauigkeitsquote (*precision)*[3] des Programms verändern. Zu Beginn des Programms definiert der Benutzer diese beiden Werte. Höhere Schwellenwerte verringern die Zahl der Alignierungen, d. h. das Programm aligniert nicht den ganzen Text, da ein Teil der Alignierungen nicht zuverlässig genug wären. Die Korrektheit der tatsächlich ausgeführten Alignierungen nimmt zu.

4.3 Methode von Church und Gale

Dieses Satzalignierungsverfahren[4] basiert auf einem einfachen statistischen Modell von Satzlängen. Es beruht auf zwei Überlegungen:

- Lange Sätze in einem Text werden i. a. durch lange Sätze übersetzt und kurze Sätze werden i. a. in der Übersetzung durch kurze Sätze wiedergegeben.
- Wenn zwei Sätze im Quell- und im Zieltext miteinander korrespondieren (d. h. die Übersetzung voneinander sind), so wirkt sich dies auf die Kombinationsmöglichkeiten der benachbarten Sätze beschränkend aus.

Für jede Satzkombination aus dem Quell- und dem Zieltext wird ein Wahrscheinlichkeitswert (Distanzmaß) berechnet, der auf der skalierten Differenz der Längen der beiden Sätze (in Buchstaben) und der Varianz dieser Differenz basiert (siehe Abschnitt 4.3.1 Berechnung des Distanzmaßes, Seite 44). Dieses Distanzmaß wird in einem sog. **dynamischen Algorithmus** (*dynamic programming framework*) zur Berechnung der wahrscheinlichsten Satzalignierungen verwendet.

Dynamische Verfahren werden oft verwendet, wenn man innerhalb zweier korrespondierender Symbolfolgen die Elemente bestimmen möchte, die einander entsprechen. Dies ist z. B. erforderlich bei Vergleichen von genetischen Code-Sequenzen verschiedener Spezies, bei Vergleichen von Sprachsignalsequenzen verschiedener Sprecher oder bei Vergleichen geologischer Schichten an unterschiedlichen Orten. Die Aufgabe des Algorithmus besteht darin, die günstigsten Zuordnungen zwischen den einzelnen Elementen zu ermitteln, so daß die Summe der sich daraus ergebenden einzelnen Distanzen minimal ist.

1. Die minimale Anzahl der Vorkommen eines Worts *x* ist *y*: *x* kommt mindestens *y* mal im Text vor.
2. Der “recall” ist hier der Quotient aus den vom Programm ausgeführten Satzalignierungen und den tatsächlich im Text vorkommenden Satzalignierungen, die dt. Übersetzung von “recall” stammt aus Henzler 1992, S. 165.
3. Die “precision” ist der Quotient aus der Zahl der korrekten Satzalignierungen und der Zahl der vom Programm ausgeführten Alignierungen, die dt. Übersetzung von “precision” stammt aus Henzler 1992, S. 165.
4. Gale & Chruch 1991.

Solange die Anordnungsreihenfolge der korrespondierenden Sätze in den beiden Texten nicht zu stark divergiert, erweist sich diese dynamische Zuordnungstechnik für Sätze als sehr gut.

4.3.1 Berechnung des Distanzmaßes

Das Distanzmaß *dist* wird durch folgende Formel berechnet:

$$dist = -\log\left(Prob\left(match|\,\delta\right)\right)$$

δ ist abhängig von den Satzlängen l_1 und l_2 und wird folgendermaßen berechnet:

$$\delta = \frac{l_2 - l_1 c}{\sqrt{l_1 s^2}}$$

> Dieses Distanzmaß basiert auf der Annahme, daß bei einer Übersetzung jedes Zeichen (Buchstabe) in einer Sprache L_1 eine Zufallszahl von Zeichen in L_2 erzeugt. Diese Zufallsvariablen seien unabhängig und nach der Normalverteilung gleichverteilt.
>
> Es sei c die mittlere erwartete Anzahl von Zeichen in L_2 für jedes Zeichen in L_1.
>
> Es sei s^2 die Varianz für die Anzahl der Zeichen in L_2 pro Zeichen in L_1.
>
> Die Parameter c und s^2 wurden empirisch aus dem engl.- dt.- frz. *UBS-Korpus*[1] bestimmt[2].
>
> dt.-engl-: c = 1,1037, d. h. c ~ 1,1; frz.-engl.: c = 1,0562, d. h. c ~ 1,06
>
> Zur Vereinfachung setzten Church und Gale c ~1.
>
> Es sei s^2 proportional zur Länge. Die Differenz zwischen dem engl.-dt., dem frz.-dt.und dem engl.-frz. Wert ist so gering, daß Church und Gale die einfachere, sprachunabhängige Abschätzung s^2~ 6,8 verwendeten.

Die Berechnung der Formel erfordert einige Umformungen. Nach dem *Bayes*'schen Theorem gilt:

1. Wirtschaftsberichte der *Union Bank of Switzerland* (UBS), ca. 15000 Wörter pro Sprache (Gale & Church 1991, S. 179).

2. Gale & Church 1991, S. 180.

$$Prob(match|\delta) = K \cdot Prob(\delta|match) \cdot Prob(match)$$

K ist eine Konstante, die für alle zu überprüfenden Kombinationen gleich ist. Sie kann daher ignoriert werden.

Die bedingte Wahrscheinlichkeit Prob(δ|match) kann folgendermaßen umgeformt werden:

$$Prob(\delta|match) = 2(1 - Prob(|\delta|))$$

Prob(|δ|*)* wird durch Integration der Standardnormalverteilung (mit Mittel = 0 und Varianz = 1) berechnet.

Prob(match) wurde empirisch bestimmt[1]. Es berücksichtigt, wie oft durchschnittlich ein Satz in einer Sprache durch genau einen Satz in der der anderen Sprache übersetzt wurde (1:1-Entsprechung), wie oft ein Satz in einer Sprache keine Entsprechung in der anderen Sprache hat oder weggelassen wurde (1:0-Entsprechung bzw. 0:1-Entsprechung), wie oft zwei Sätze in einer Sprache nur einem Satz in der anderen Sprache entsprechen (2:1-Entsprechung bzw. 1:2-Entsprechung), wie oft zwei Sätze in einer Sprache zwei Sätzen in der anderen Sprache entsprechen (2:2-Entsprechung).

Tabelle 4. Satzentsprechungen und Wahrscheinlichkeiten aus dem dt.-engl.-frz. UBS-Korpus

Art der Entsprechung	Häufigkeit	Prob(match)
1:1	1167	0,89
1:0 bzw. 0:1	13	0,0099
2:1 bzw. 1:2	117	0,089
2:2	15	0,0011

1. Gale & Church 1991, S. 180.

4.3.2 Berechnung der Distanzfunktion

Die Distanzfunktion *two_side_distances* berechnet aufgrund der vier Parameter x_1, y_1, x_2, y_2[1] die jeweilige Distanz nach der angegebenen Formel (siehe Abschnitt 4.3.1 Berechnung des Distanzmaßes, Seite 44) und berücksichtigt dabei sechs verschiedene Kombinationsmöglichkeiten.

1. *two_side_distances* (x_1, y_1; 0, 0)

   1:1-Entsprechung: x_1 ist durch y_1 übersetzt. (Substitution */ substitution*)

2. *two_side_distances* (x_1, 0; 0, 0)

   1:0-Entsprechung: x_1 ist nicht übersetzt. (Löschung */ deletion*)

3. *two_side_distances* (0, y_1; 0, 0)

   0:1-Entsprechung; y_1 ist eingefügt worden. (Einfügung / *insertion*)

4. *two_side_distances* (x_1, y_1; x_2, 0)

   2:1-Entsprechung: x_1 und x_2 sind durch y_1 übersetzt. (Verkürzung / *contraction*)

5. *two_side_distances* (x_1, y_1; 0, y_2)

   1:2-Entsprechung: x_1 ist durch y_1 und y_2 übersetzt. (Erweiterung / *expanding*)

6. *two_side_distances* (x_1, y_1; x_2, y_2)

   2:2-Entsprechung: x_1 und x_2 sind durch y_1 und y_2 übersetzt, evtl. in vertauschter Reihenfolge der Sätze. (Verschmelzung / *merging*).

4.3.3 Algorithmus

Das Satzalignierungsprogramm besteht aus zwei Arbeitsschritten. Zuerst werden Abschnitte aligniert, und dann werden die Sätze innerhalb eines Abschnitts aligniert.

Der dynamische Algorithmus arbeitet rekursiv: Er nimmt Satzpaare als Eingabe und berechnet ihre Distanzmaße. Um eine mögliche Entsprechung zwischen einem Satz s_i und seiner Übersetzung t_j zu finden, berechnet das Programm rekursiv ihre minimale Distanz D(i, j) unter Berücksichtigung der sechs Arten von Entsprechungen.

Die folgende Rechenvorschrift beschreibt den dynamischen Algorithmus:

s_i (i=1 ... I) seien die Sätze in Sprache L_1 in Text 1,

1. x_1, x_2: Sätze des Quelltexts bzw. deren Satzlängen; y_1, y_2: Sätze des Zieltexts bzw. deren Satzlängen..

t_j (j=1 ... J) seien die Sätze in Sprache L_2 in Text 2.

d steht für die Distanzfunktion (*two_side_distances*).

D (i, j) sei die Summe der Distanzen zwischen den Sätzen s_1 ... s_i und ihren Übersetzungen t_1 ... t_j, wobei alle Kombinationen rekursiv gebildet werden, so daß die Summe minimal wird.

Rechenvorschrift:

$$D(i,j) = min \left\{ \begin{array}{c} D(i,j-1) + d(0, t_j; 0, 0) \\ D(i-1,j) + d(s_i, 0; 0, 0) \\ D(i-1,j-1) + d(s_i, t_j; 0, 0) \\ D(i-1,j-2) + d(s_i, t_j; 0, t_{j-1}) \\ D(i-2,j-1) + d(s_i, t_j; s_{i-1}, 0) \\ D(i-2,j-2) + d(s_i, t_j; s_{i-1}, t_{j-1}) \end{array} \right\}$$

Die beiden beschriebenen Algorithmen wurden implementiert, getestet und sowohl einzeln als auch komparativ evaluiert. Für den Test wurden zweisprachige Texte aus dem EPA-Korpus mit den Sprachpaaren Englisch-Deutsch und Französisch-Deutsch verwendet.

4.3.4 Implementierung

Der Algorithmus wurde in C implementiert und mit verschiedenen Textkorpora getestet.

Der im Anhang von Gale und Church (1993) veröffentlichte C-Code diente als Grundlage für die Implementierung. Das von mir erstellte Programm hat eine ähnliche Struktur und verwendet die gleichen Grundfunktionen für den rekursiven Alignierungsprozeß und die Berechnung des Distanzmaßes. Die vorhergehende Textanalyse (Bestimmung der Satzlängen usw.) wurde jedoch anders programmiert.

4.3.4.1 Programmstruktur

Zu Beginn bestimmt das Programm die Anzahl der Abschnittsmarkierungen. Wenn diese Zahl unterschiedlich für die beiden Texte ausfällt, terminiert das Programm ohne

Alignierung[1], andernfalls erfolgt die Bestimmung der Anzahl der Sätze und der Satzlängen.

Damit stehen alle für die Alignierung notwendigen Informationen zur Verfügung. Die Funktion *seq_align* verwendet fünf Parameter zur iterativen Alignierung:

x*:* Array der Satzlängen aus Text 1

y: Array der Satzlängen aus Text 2

nx: Zahl der Sätze pro Abschnitt in Text 1

ny: Zahl der Sätze pro Abschnitt in Text 2

dist_func: vordefinierte Distanzfunktion.

Die Funktion *seq_align* erzeugt eine Satzalignierungstabelle. Eine Funktion gibt dann die Satzpaare, die diesen Alignierungen entsprechen, gemäß der Tabelle aus, wobei zusätzlich für jedes Satzpaar auch die berechneten Distanzmaße angegeben werden.

4.3.5 Erweiterung der Church-Gale-Methode

In der Arbeit von Gaussier (1995) wurden engl. und frz. Texte für die Extraktion von Terminologie aligniert.

Gaussier nutzt *transwords*[2], d. h. Wörter, die in ähnlicher Form in den engl. und den frz. Texten vorkommen. Dazu gehören z. B. Zahlen, Eigennamen und Formatierungsinformation. Dieses Konzept ist inspiriert von der Definition der *cognates* von Simard et al. (1992). *Cognates* sind Wörter, deren erste vier Buchstaben identisch sind, z. B. *financed* und *financier.* Diese Definition erfaßt aber einerseits Wortpaare wie *government* und *gouvernement* nicht, obwohl eine Verbindung zwischen ihnen besteht, und schließt andererseits falsche Wortpaare wie *international* (engl.) und *intercontinental* (frz.) ein. Die von Gaussier verwendete Definition der *transwords* berücksichtigt diese Probleme, es wird aber nicht erklärt, durch welche Kriterien dies erfolgt. Seine Untersuchung zeigt, daß alignierte Satzpaare eine weit höhere Zahl von *transwords* aufweisen als nicht alignierte Satzpaare. Gaussier merkt allerdings an, daß das Prinzip der *cognates* nur bei Sprachen interessant ist, die miteinander verwandt sind.

Der Ähnlichkeitswert eines Satzpaares berücksichtigt drei Parameter:

- die Wahrscheinlichkeit des Entsprechungstyps (1:1-Entsprechung, 1:2-Entsprechung usw.)
- die Satzlängen
- die Anzahl der *transwords*.

1. Church und Gale haben für die Alignierung solcher Texte (*noisy texts*) einen anderen Algorithmus konzipiert (Church & Gale 1993).

2. Gaussier 1995, S. 26.

Ein dynamisches Programmierverfahren (Viterbi-Algorithmus) berechnet aus diesen Werten die *optimalen* Satzkombinationen. Die Korrektheit der Alignierung (Fachtexte aus dem Bereich der Telekommunikation) liegt bei 98 %.

4.4 Evaluierung

4.4.1 Vorbereitung des Testmaterials

4.4.1.1 Satzendeerkennung

Es wurden drei Programme (in Perl) zur einzelsprachlichen Satzendeerkennung (engl., dt., frz) geschrieben. Diese Programme haben zwei Aufgaben:

1. Punkte erkennen, die kein Satzende markieren.
2. Einheiten erkennen und markieren, die nicht mit einem Punkt abgeschlossen sind, aber von den Alignierungsprogrammen wie Sätze behandelt werden sollen.

Punkte, die kein Satzende markieren, treten in folgenden Fällen auf:

- zur Strukturierung (mit römischen und arabischen Zahlen): *1. 1, V.*
- in Datumsangaben: *am 20. Dezember 1978, die Fassung vom 7. 7. 1983, version of 7. 7. 1989, version du 7.7. 1989*
- in Abkürzungen.

Ein Teil der Abkürzungen ist einzelsprachlich, der andere Teil ist international (z. B. bei Maßeinheiten) oder textsortenspezifisch. Es kommen jedoch in den Texten - beispielsweise in Zitaten - auch Abkürzungen vor, die eigentlich in der Sprache des Textes nicht existieren (z. B.: *chambre de rec*. für *chambre de recours* in einem dt. Text). Dieselbe Abkürzung kann auch in Variationen auftreten (*d. h*. und *d.h.*, *Abl*., *ABl*. und *ABL*. als Abkürzung für *Amtsblatt*).

Manche Einheiten, beispielsweise Titel, Überschriften und Deskriptoren, sind nicht mit einem Punkt abgeschlossen, können aber durch die Annotation des Korpus leicht automatisch lokalisiert und so markiert werden, daß sie von den Alignierungsprogrammen wie Sätze behandelt werden.

4.4.1.2 Markierung von Abschnittsgrenzen

Alle Dokumente des Testkorpus sind sehr gut strukturiert und entsprechend ihrer Struktur annotiert (siehe Abschnitt 3.3.1 Strukturierung eines Dokuments, Seite 28). Die Dokumente aus dem EBK-Korpus unterscheiden sich zwar im einzelnen in ihrem Aufbau, bestimmte Struktureinheiten kommen aber in jedem Text der Entscheidungen der Beschwerdekammer vor. So wurden in jedem Dokument automatisch vier Teile lokalisiert, die *Abschnitte* genannt werden.

1. **Abschnitt 1**: vom Beginn des Dokuments bis zur Annotation <FSU>
2. **Abschnitt 2**: von der Annotation <FSU> bis zur Annotation <RES>
3. **Abschnitt 3**: von der Annotation <RES> bis zur Annotation <ORD>
4. **Abschnitt 4**: von der Annotation <ORD> bis zum Ende des Dokuments.

So ist sichergestellt, daß jedes für die Alignierung verwendete Dokument die gleiche Anzahl und Struktur von Abschnitten aufweist, allerdings variiert die Länge der einzelnen Anschnitte dadurch erheblich. Durchschnittlich liegt in einem Dokument der Anteil von Abschnitt 1 bei 10 %, Abschnitt 2 liegt bei 25 %, Abschnitt 3 bei 63 % und Abschnitt 4 liegt bei nur 2 %. In langen Dokumenten des untersuchten Korpus enthielt der dritte Abschnitt bis zu 165 Sätze. Zu lange Abschnitte erwiesen sich in Einzelfällen als problematisch für die Alignierung nach dem Church-Gale-Algorithmus.

4.4.1.3 Bestimmung der Parameter für den Church-Gale-Algorithmus

4.4.1.3.1 Bestimmung von c

Der Parameter c wurde für einen Teilkorpus des EPA-Korpus (ca. 20 000 Wörter pro Sprache) berechnet. Die Werte entsprechen in etwa den von Church und Gale für das UBS-Korpus ermittelten Werten.

dt.-engl.: c = 1,0381

dt.-frz.: c = 1,0490

engl.-frz.: 1,0890.

4.4.1.3.2 Bestimmung von Prob(match)

Die Häufigkeiten der Satzentsprechungen wurden für die Sprachpaare dt.-engl. und dt.-frz. bestimmt. Dafür wurde in einem alignierten Testkorpus untersucht, welche Art von Entsprechungen unter den korrekten Satzkombinationen vorkommen.

Satzentsprechungen im dt.-engl. Testkorpus

In einem Testkorpus von 27 000 Wörtern kam es zu folgender Verteilung:

Tabelle 5. Satzentsprechungen und Wahrscheinlichkeit im dt.-engl. Testkorpus

Art der Entsprechung	Häufigkeit	Prob(match)
1:1	924	0,956
1:0 bzw. 0:1	0	0,0
2:1 bzw. 1:2	40	0,041
2:2	2	0,002

Satzentsprechungen im dt.-frz. Testkorpus

In einem Testkorpus von ca. 40 000 Wörtern kam es zu folgender Verteilung:

Tabelle 6. Satzentsprechungen und Wahrscheinlichkeit im dt. -frz. Testkorpus

Art der Entsprechung	Häufigkeit	Prob(match)
1:1	1113	0,94
1:0 bzw. 0:1	0	0,0
2:1 bzw. 1:2	67	0,056
2:2	2	0,0016

Beispiele für die verschiedenen Arten von Entsprechungen

Tabelle 7. Beispiel für 1:1-Entsprechung

frz. Text	dt. Text
20: <FSU> b) La titulaire du brevet a en particulier, dans les exemples cités dans le brevet en litige, abaissé de toute une puissance de dix, par rapport aux chiffres cités dans la demande antérieure, les valeurs relatives à la teneur en sodium retenues dans le document sur lequel se fonde la priorité.	22: <FSU> b) Insbesondere habe die Patentinhaberin die prioritätstragenden Natriumwerte in den Beispielen des Streitpatents gegenüber denen der Voranmeldung um eine ganze Zehnerpotenz nach unten verschoben.

Tabelle 8. Beispiel für 2:1-Entsprechung

frz. Text	dt. Text
21: Il ne s'agit pas là d'une rectification d'erreurs de frappe, mais d'une correction importante apportée au contenu technique de la demande antérieure, qui n'est pas fondée sur l'exposé de celle-ci; dans le domaine de la chimie, les exemples de réalisation de l'invention fournis à l'appui de la revendication de l'objet du brevet font en effet toujours partie de la demande; or les exemples fournis en l'occurrence sont inexacts et ne peuvent être corrigés a posteriori, ce qui devrait normalement entraîner la perte du droit de priorité revendiqué pour le brevet en litige.	23: Hierbei handle es sich nicht um die Korrektur von Schreibfehlern, sondern um eine durch die Offenbarung aus der Voranmeldung nicht gestützte wesentliche Korrektur eines technischen Sachverhalts; denn zu einer chemischen Anmeldung gehören stets die den Patentgegenstand belegenden Ausführungsbeispiele, die eben hier unrichtig und nicht nachträglich korrigierbar seien. 24: All dies müsse den Prioritätsverlust für das Streitpatent nach sich ziehen.

Tabelle 9. Beispiel für 2:2-Entsprechung

frz. Text	dt. Text
18: <FSU> a) un brevet ne peut être considéré comme valable si l'enseignement essentiel de ses revendications est en contradiction avec la description s'y rapportant. 19: La revendication du brevet exige en l'occurrence l'absence d'alcali, alors que les exemples montrent au contraire que l'on opère en présence d'alcali, ce que confirme également le reste de la description, où il est recommandé d'utiliser des produits chimiques pauvres en sodium, donc contenant du sodium, pour la préparation des zéolites, et constaté que des traces de sodium accélèrent la cristallisation.	20: <FSU> a) Ein Patent könne keinen Bestand haben, dessen anspruchsgemäße Kernaussage im Widerspruch zu der zugehörigen Patentbeschreibung stehe, indem der Patentanspruch die Abwesenheit von Alkali fordere, die Beispiele aber gerade das Arbeiten in Gegenwart von Alkali belegen. 21: All dies werde auch durch die übrige Beschreibung bestätigt, wonach die Verwendung natriumarmer, das heißt natriumhaltiger Chemikalien für die Zeolitherstellung empfohlen und festgestellt werde, daß Natriumspuren die Kristallisation beschleunigen.

Echte Nicht-Übersetzungen von Sätzen, d. h. 1:0-Entsprechungen, kamen im Korpus nicht vor. Das folgende Beispiel stammt aus einem Dokument, dessen Originalsprache Englisch ist. Die Liste der Deskriptoren (beginnend mit “<KEY>”) wurde in der dt. Übersetzung beibehalten. In der frz. Übersetzung wurde ein Deskriptor weggelassen,

wahrscheinlich, weil er in einem anderen komplexeren Deskriptor, der kurz darauf folgt, schon enthalten ist.

Tabelle 10. Beispiel für 1:0-Entsprechung

frz. Text	dt. Text
	18: <KEY> Automatischer Widerruf
18: <KEY> Maintien du brevet sous une forme modifiée	19: <KEY> Patent in geändertem Umfang aufrechterhalten
19: <KEY> Non-respect des délais prévus pour acquitter la taxe d'impression et pour produire les traductions.	20: <KEY> Druckkostengebühr und Übersetzung nicht fristgerecht entrichtet bzw. eingereicht.
20: <KEY> Révocation automatique du brevet dès l'expiration de ces délais	21: <KEY> Widerruf des Patents - sofortiger automatischer bei Fristablauf

4.4.2 Testreihen und Ergebnisse

Es wurden vier Testreihen durchgeführt, um die Performanz der beiden Algorithmen in bezug auf verschiedene Kriterien zu testen. Dazu gehören:

- Art des Textes
- Länge des Textes
- Sprachpaar: dt.-engl., dt.-frz. und engl.-dt.

Bestimmung der Performanz

Die *Korrektheit* des Church-Gale-Algorithmus berechnet sich aus dem Quotienten der korrekten Satzalignierungen und der Gesamtzahl der ausgeführten Satzalignierungen, denn dieser Algorithmus aligniert immer alle Sätze, d. h. die *coverage* beträgt immer 100 %.

Die Bewertung des Kay-Röscheisen-Algorithmus berücksichtigt die Anzahl der korrekten Satzalignierungen in der SAT (*relative Korrektheit*) und den Quotienten der korrekten Satzalignierungen und der tatsächlich im Text existierenden Satzalignierungen (*absolute Korrektheit*) (siehe Abschnitt 4.2.4 Evaluierung, Seite 41). Besteht ein Testkorpus aus mehreren Texten, so sind die Angaben zur Korrektheit und Abdeckung immer Durchschnittswerte.

Fehlerhafte Alignierungen

Fehlerhafte Satzalignierungen sind entweder vollkommen **falsch** oder **partiell inkorrekt**. Eine Satzalignierung ist falsch, wenn die kombinierten quell- und zielsprachlichen Sätze in keinem Fall die Übersetzung voneinander sind. Eine Satzalignierung ist partiell inkorrekt, wenn eine Teilübereinstimmung - der ganze Satz oder ein Nebensatz - vorliegt. Partiell inkorrekte Übersetzungen kommen des öfteren dann vor, wenn ein Algorithmus eine 2:1-Entsprechung berechnet hat. Sie werden mit dem Faktor 0,5 bewertet, falsche Alignierungen dagegen mit dem Faktor 1.

Textlängenangaben bei Korpora

Wird die Textlänge durch die Anzahl der Wörter ausgedrückt, so ist meist das frz. Dokument das längste, das entsprechende dt. Dokument das kürzeste und das entsprechende engl. Dokument liegt zwischen den beiden Werten. Die Längenangaben (Anzahl der Wörter) bilingualer bzw. trilingualer Texte in dieser Arbeit beziehen sich immer auf die mittlere Länge.

4.4.2.1 Testreihe 1: Vergleichstest mit Gesetzestext, drei Sprachpaare, mittlere - längere Texte

Testkorpus: Europäisches Patentübereinkommen (EPÜ), S. 24-63, engl.-dt.-frz., ca. 5000 Wörter, ca. 300 Sätze.

Tabelle 11. Ergebnis Testreihe 1: Vergleichstest, Gesetzestext, mittlere - längere Texte

Sprachpaar	Church-Gale-Algorithmus Korrektheit	Kay-Röscheisen-Algorithmus: absolute Korrektheit	Kay-Röscheisen-Algorithmus: relative Korrektheit	Kay-Röscheisen-Algorithmus: *coverage*
dt.-engl.	100 %	100 %	100 %	100 %
dt.-frz.	100 %	92,45 %	92,45 %	100 %
engl.-frz.	100 %	94, 34 %	94,34 %	100 %

4.4.2.2 Testreihe 2: Vergleichstest mit Dokumenten aus dem EBK-Korpus, engl.-dt., längere Texte

Testkorpus: 4 Dokumente aus dem EBK-Korpus, Sprachpaar engl.-dt., ca. 27 000 Wörter, durchschnittliche Satzzahl pro Dokument: ca. 250 Sätze.

Tabelle 12. Ergebnis Testreihe 2: Vergleichstest, EBK-Korpus, längere Texte

Sprachpaar	Church-Gale-Algorithmus Korrektheit	Kay-Röscheisen-Algorithmus: absolute Korrektheit	Kay-Röscheisen-Algorithmus: relative Korrektheit	Kay-Röscheisen-Algorithmus: *coverage*
engl.-dt.	96, 71 %	86,13 %	89,57 %	88, 45 %

4.4.2.3 Testreihe 3: Vergleichstest mit Dokumenten aus dem EBK-Korpus, engl.-dt., kürzere Texte

Testkorpus: 11 Dokumente aus dem EBK-Korpus, Sprachpaar engl.-dt., ca. 20 500 Wörter, d. h. durchschnittlich 1870 Wörter und 74 Sätze pro Dokument, kürzestes Dokument: 40 Sätze, längstes Dokument: 120 Sätze.

Tabelle 13. Ergebnis Testreihe 3: Vergleichstest, EBK-Korpus, kürzere Texte

Sprachpaar	Church-Gale-Algorithmus Korrektheit	Kay-Röscheisen-Algorithmus: absolute Korrektheit	Kay-Röscheisen-Algorithmus: relative Korrektheit	Kay-Röscheisen-Algorithmus: *coverage*
engl.-dt.	92,50 %	62,65 %	89,75 %	70,45 %

4.4.2.4 Testreihe 4: Test des Church-Gale-Algorithmus mit Dokumenten aus dem EBK-Korpus, Sprachpaar dt.-frz.

Testkorpus: 6 Dokumente aus dem EBK-Korpus, Sprachpaar dt.-frz., ca. 38 000 Wörter, kürzere und längere Texte, kürzestes Dokument: 140 Sätze, längstes Dokument: 270 Sätze.

Tabelle 14. Testreihe 4: Church-Gale-Algorithmus, EBK-Korpus, kürzere und längere Texte

Sprachpaar	Church-Gale-Algorithmus Korrektheit
dt.-frz.	97,12 %

4.4.3 Vergleichende Evaluierung

4.4.3.1 Bewertung der Testreihen

Die Testreihen sollten nicht nur die Leistungsfähigkeit der beiden Algorithmen vergleichen, sondern auch Kriterien liefern, welche Methode für welche Texte die besseren Ergebnisse liefert.

In der Testreihe 1 wurde ein Ausschnitt aus der Europäischen Patentübereinkommen (EPÜ) aligniert. Die Texte des EPÜ sind Vorschriften für die Erteilung europäischer Patente und haben somit "Gesetzescharakter". Sie sind ohne jede Abweichung, quasi "1:1" übersetzt und deshalb sehr leicht zu alignieren, was sich in den sehr guten Ergebnissen beider Algorithmen zeigt.

In der Testreihe 2 wurden längere Texte aus dem EBK-Korpus des Sprachpaars dt.-engl. mit beiden Algorithmen aligniert. Die Korrektheit des Church-Gale-Algorithmus liegt hier ca. 10 % höher als die des Kay-Röscheisen-Algorithmus.

In der Testreihe 3 wurden vorwiegend kürzere Texte des Sprachpaars dt.-engl. aus dem EBK-Korpus mit beiden Algorithmen aligniert. Die Ergebnisse bestätigen die Evaluirung von Wallner (1994): bei kürzeren Texten arbeitet der Kay-Röscheisen-Algorithmus wesentlich schlechter, d. h. die absolute Korrektheit liegt nur bei 62 %.

In der Testreihe 4 sollte überprüft werden, ob die guten Ergebnisse des Church-Gale-Algorithmus auch für das Sprachpaar dt.-frz. gelten. Die Vermutung, daß die früheren Ergebnisse sprachunabhängig waren, bestätigte sich. Die Korrektheit lag mit 97 % sogar noch höher als für die dt.-engl. Texte, was z. T. auf eine bessere Satzendeerkennung zurückzuführen ist.

Alle für die Evaluierung verwendeten Texte waren **leicht zu alignieren**. Es müßte überprüft werden, inwieweit die erzielten Ergebnisse auch für die Alignierung anderer Textsorten (z. B. allgemeinsprachliche Zeitungsartikel, technische Dokumentation usw.) gültig sind.

Geschwindigkeit

Bei einer Textlänge von 150 Sätzen arbeitet das Church-Gale-Programm drei bis viermal schneller als die Prolog-Implementierung der Kay-Röscheisen-Methode. Dieser Unterschied ist auch darauf zurückzuführen, daß die beiden Algorithmen nicht in derselben Programmiersprache implementiert wurden. Wären beide Programme in C geschrieben, wäre der Church-Gale-Algorithmus wahrscheinlich nur zweimal schneller. Die Ausführungszeit des Kay-Röscheisen-Algorithmus ist abhängig von der Zahl der Iterationen, die sich je nach den vom Benutzer definierten Schwellenwerten ändert.

4.4.3.2 Kay-Röscheisen-Algorithmus: Bemerkungen

- Die Korrektheit der Alignierung steigt mit zunehmender **Textlänge** (mehr als 250 Sätze).
- Die **Abdeckung** (*coverage*) erreicht nie 100%, da der Algorithmus nur die besten Alignierungen berücksichtigt.
- Der Algorithmus nimmt **keine Plausibilitätsüberprüfung der Satzlänge** vor. Einige der ausgegebenen Alignierungen sind unwahrscheinlich, da die Längen der betreffenden Sätze zu stark differieren. Solche Alignierungen sollten ausgeschlossen werden. Außerdem sollte die Zahl der Wortpaare, die für ein aligniertes Satzpaar stehen, proportional zur Satzlänge sein, d. h. die Alignierung längerer Sätze sollte sich auf mehr Wortpaare stützen als die Alignierung kürzerer Sätze.
- Vordefinierte grammatische **Funktionswörter** werden aufgrund ihrer Frequenz nicht für die Wortalignierung herangezogen. Eine größere Menge von vordefinierten sprachspezifischen (und evtl. auch textspezifischen) Funktionswörtern erhöht die Korrektheit beträchtlich. Im EPA-Korpus sind manche spezielle juristische und technische Abkürzungen so frequent wie gewöhnliche Funktionswörter. Diese *speziellen* Funktionswörter sollten nicht für die Wortalignierung herangezogen werden.
- Die Bestimmung zusätzlicher **Ankerpunkte** (Zahlen, Datumsangaben, Eigennamen etc.) erhöht die Korrektheit.
- Die Segmentierung von **Komposita** für Sprachen wie das Deutsche verbessert die Qualität der Wortalignierung. In einem kleinen Korpus (zwei deutsche Dokumente) wurden die Komposita automatisch segmentiert[1] und manuell überprüft. Die Korrektheit der WAT erhöhte sich um fast 40 %. Das entgegengesetzte Verfahren sollte auch getestet werden: Lokalisierung von engl. bzw. frz. Komposita vor der Alignierung mit einem dt. Text und Evaluierung der resultierenden WAT's.

4.4.3.3 Church-Gale-Algorithmus: Bemerkungen

- Die **Textlänge** hat keinen direkten Einfluß auf die Korrektheit der Alignierung, unter der Voraussetzung, daß längere Texte in Absätze unterteilt sind. In diesem Fall erzielt der Algorithmus ungefähr die gleichen Ergebnisse für kürzere und längere Texte.
- Die **Zahl der Absätze** in den beiden Texten muß übereinstimmen, sonst wird keine Alignierung vorgenommen.
- Der Algorithmus ist nur **satzlängenbasiert** und nicht "intelligent", d. h. wenn einmal zwei Sätze falsch aligniert wurden, besteht die Gefahr, daß die restlichen Sätze des Abschnitts auch falsch aligniert werden, wenn sie "passende" Satzlängen haben. Oft kann der Algorithmus vor dem Ende des Abschnitts keinen korrekten Wiederanknüpfungspunkt finden.

1. Vgl. dazu Maier-Meyer 1995, S. 100.

- Der Algorithmus erfordert eine **möglichst korrekte Satzendeerkennung**. Eine Analyse der fehlerhaften Alignierungen zeigte, daß etwa 30 % der Fehler auf eine inkorrekte Satzendeerkennung zurückzuführen sind.
- Da die **Korrektheit mit steigender Abschnittslänge abnimmt**, sollten zumindest in längeren Absätzen zusätzliche Plausibilitätstest für Alignierungen durchgeführt werden, z. B. Ankerpunkte wie Zahlen, Eigennamen usw. Der längste Abschnitt im Testkorpus umfaßte 150 Sätze, damit wurden aber noch passable Ergebnisse erzielt. Diese Beobachtung kann aber auch auf die Tatsache zurückgeführt werden, daß das Testkorpus nicht sehr schwierig zu alignieren war.
- Das **Distanzmaß** ist nicht immer ein Hinweis für die Erkennung inkorrekter Alignierungen. Gale und Church[1] berichten, daß das Distanzmaß ein Prädiktor für die Qualität einer Alignierung ist. Diese Behauptung wurde aber durch das Testkorpus nicht bestätigt. Es wurde sowohl festgestellt, daß Alignierungen mit einem hohen Distanzmaß in manchen Fällen korrekt waren, als auch, daß Alignierungen mit einem niedrigen Distanzmaß falsch waren. Es ist also nicht möglich, durch Festlegung eines Schwellenwerts inkorrekte Alignierungen maschinell zu erkennen.

Distanzmaß

Für ein dt.-frz. Teilkorpus (ca. 25 000 Wörter pro Sprache) wurden die Distanzen für die korrekten Alignierungen analysiert.

Tabelle 15. Distanzmaße der Alignierungen (Verteilungen)

Distanzmaß (Größe)	Frequenz
Distanz liegt zwischen 0 und 100	74 %
Distanz liegt zwischen 100 und 200	16 %
Distanz liegt zwischen 100 und 300	6 %
Distanz liegt zwischen 300 und 500	2,5 %
Distanz größer 500	1,5 %

4.5 Anwendungen der Alignierung

Zweisprachige Korpora enthalten Informationen, die für verschiedene Zwecke gebraucht werden und die manuell zusammengestellt werden müßten, falls keine Alignierungsverfahren zur Verfügung ständen. Zu den Anwendungen, die Informationen aus zweisprachigen Korpora verwenden, gehören z. B.

- Tools für die Nutzung bilingualer Korpora (Warwick et al. 1993, Gale und Church 1991)

1. Gale & Church 1991, S. 46.

- Entsprechungen von allgemeinsprachlichen Nominalphrasen in bilingualen Korpora (Kupiec 1993)
- Entsprechungen von fachsprachlichen Nominalphrasen (siehe Abschnitt 2.2.2 Das Programm Termight, Seite 13; Abschnitt 2.2.3 Der Ansatz von Gaussier, Seite 15 und Abschnitt 2.2.4 Der Ansatz von Eijk, Seite 17)
- Lesartendisambiguierung (Gale et al. 1992, Dagan und Itai 1993)
- statistisch-basierte maschinelle Übersetzung (Brown et al. 1990, Sato und Nagao 1990)
- Übersetzungsspeicher (siehe Abschnitt 2.2.5 Das EURAMIS-Projekt der Europäischen Gemeinschaft, Seite 19)
- multilinguales Information Retrieval (Landauer und Littman 1990).

4.5.1 Lesartendisambiguierung

Die Grundidee besteht darin, lexikalische Ambiguitäten in einer Sprache durch die Verwendung von Daten über die lexikalischen Relationen in einer anderen Sprache zu lösen.

Gale et al.[1] erzielen eine Korrektheit von 90 % bei der Unterscheidung der verschiedenen Lesarten eines englischen Substantivs wie *sentence,* das im Franz. durch *peine* ("Strafe") oder *phrase* ("Satz") übersetzt wird. In der Trainingsphase werden Vorkommen des Wortes in den unterschiedlichen Lesarten gesammelt. In der Testphase erhält der Algorithmus ein neues Vorkommen des Wortes als Eingabe (noch nicht analysierte Textstelle) und versucht, dieses Vorkommen einer der beiden Lesarten zuzuordnen. Diese Aufgabe wird dadurch gelöst, daß der Kontext der zu analysierenden Textstelle mit bereits analysierten Textstellen verglichen wird. Der Vergleich benützt einen *Bayes*'schen Parameter, der schon in ähnlichen Anwendungen, z. B. Autoridentifizierung[2] oder im Information Retrieval[3] erfolgreich angewandt worden war.

Dagan und Itai[4] gehen anders vor: sie verwenden Parsing und ein zweisprachiges Wörterbuch. Syntaktische Relationen zwischen Wörtern werden mit einem quellsprachlichen Parser identifiziert und mit Hilfe eines zweisprachigen Wörterbuchs auf den zielsprachlichen Text abgebildet. Die Auswahl der präferentiellen Lesarten verwendet eine Statistik über lexikalische Relationen in der Zielsprache. Sie basiert auf einem statistischen Modell und einem *constraint-propagation*-Algorithmus.

1. Gale et al. 1992.
2. Mosteller & Wallace 1964.
3. Salton 1989.
4. Dagan & Itai 1993.

4.5.2 Statistische Methoden in der Maschinellen Übersetzung

In den neueren Ansätzen zur MÜ werden bilinguale Korpora als lexikalische Resourcen eingesetzt, indem bestehende Übersetzungen aufbereitet und als Übersetzungsmuster verwendet werden. Auf dieser Idee basieren auch Übersetzungsspeicher (*translation memories*). Die Übersetzungsmuster kann man sich als Beispielsammlungen vorstellen, die aus bilingualen alignierten Korpora extrahiert wurden. Die bekanntesten Methoden stammen von Sato und Nagao[1] und Brown et al.[2] Im folgenden soll die Methode von Sato und Nagao kurz skizziert werden.

Das Verfahren basiert auf einer großen Menge von Übersetzungsbeispielen inklusive Wortkorrespondenzen, die mit Satz- und Wortalignierungsalgorithmen generiert wurden.

Beispiel:

Der folgende Satz soll übersetzt werden:

(1) He buys a book on international politics.

Im Wissenspeicher stehen als “ähnliche” Beispiele:

(2) He *buys* a a notebook.

(2’) Er *kauft* ein Notizbuch.

(3) I read *a book on international politics*.

(3’) Ich lese ein *Buch über internationale Politik*.

Durch Zusammenfügen der Fragmente ergibt sich der Zielsatz:

(1’) Er kauft ein Buch über internationale Politik.

Vorgehensweise:

1. Der zu übersetzende Satz muß in einzelne Fragemente geteilt werden, zu denen jeweils Bespielsübersetzungen existieren. Der zerlegte Ausgangssatz wird als *matching expression* bezeichnet. Ein Ausgangssatz kann theoretisch in verschiedene Fragmente unterteilt werden.
2. Anhand der Beispielübersetzungen werden die Fragmente in die Entsprechungen der Zielsprache übertragen.
3. Die Kombination der Zielfragmente ergibt die *matching expression* des Zielsatzes.

1. Sato & Nagao 1990.
2. Brown et al. 1990.

5 Nominale Fachterminologie im Französischen

Verschiedene Disziplinen, z. B. Linguistik, Übersetzungswissenschaft, Computerlinguistik und Dokumentationswissenschaft, beschäftigen sich mit der Frage, wie im Französischen ein Fachterminus zu definieren ist. Aufgrund der Besonderheiten der frz. Wortbildung ist die Definition der Fachterminologie immer auch eng mit der Frage der Kompositabildung verflochten. Da alle Ansätze einen großen Teil der Fachterminologie als Komposita einstufen, soll zunächst ein kurzer Überblick über die Kompositaforschung gegeben werden.

Danach werden drei Arbeiten aus der Computerlinguistik vorgestellt, die sich mit der Definition, Extraktion und Evaluierung von frz. Fachterminologie beschäftigen. Die Ergebnisse dieser Arbeiten konnten z. T. in der Untersuchung des EPA-Korpus umgesetzt werden.

Eine dreigliedrige Studie beschreibt die Extraktion der Fachterminologie aus dem EPA-Korpus:

- Studie 1: Was soll extrahiert werden? (linguistische Beschreibung der Terminologie eines Teilkorpus)
- Studie 2: Wie soll extrahiert werden? (Bestimmung einer geeigneten Methode)
- Studie 3: Ergebnisse

5.1 Kompositaforschung

Die verschiedenen Forschungsansätze zu Nominalkomposita im Frz. zeigen zum einen eine enorme Vielfalt, zum anderen ist ihnen aber gemeinsam, daß kein Ansatz *nur* mit linguistischen Kriterien arbeitet. Die Erkennung von Sequenzen, die in der Allgemeinsprache als Komposita (*noms composés*) und in Fachtexten als Termini (*termes*) bezeichnet werden, ist für jede lexikographische Arbeit, im besonderen aber im Dokumentationswesen von großer Bedeutung. Ein beträchtlicher Teil der computerlinguistischen Arbeiten zu Nominalkomposita sind im Kontext des Dokumentationswesens enstanden (frz. *documentation* oder auch *informatique documentaire* genannt).

In der Diskussion über die Definition des Begriffs stehen die Wortbildungsverfahren und der Status der Komposita im Mittelpunkt.

> Nous hésitons à employer noms composés sans guillemets...
>
> “Nom composé”: une étiquette piégeante. Pour certains, il s’agit du mode de formation d’une partie des mots. Pour d’autres, cela renvoie aux dénominations complexes homologuées par la collectivité langagière. Pour d’autres encore, il s’agit des séquences nominales qui n’offrent pas toute la plasticité syntaxique du syntagme nominal. Et le débat tourne souvent au dialogue de sourds, sur l’extension de l’ensemble considéré ou sur le statut de telle ou telle suite de mots. Il en

> va de même pour deux autres des termes autour desquels se structure le débat dans le domaine: figement et lexicalisation.[1]

5.1.1 Kriterien für Komposita

5.1.1.1 Lexikalisierung

Komposita zeichnen sich meist durch eine starke Kohäsion aus. Sie wird dadurch sichtbar, daß bei Komposita z. B. Einfügungen innerhalb des Kompositums oder das Hinzufügen modifizierender Elemente weitaus stärkeren Restriktionen unterliegen als bei gewöhnlichen Nominalgruppen. M. Gross hat in seinen Arbeiten über Komposita[2] eine Reihe von Tests dargestellt, die dazu dienen, diesen Aspekt der Lexikalisierung (Übersetzung für frz. *figement*) zu beurteilen. Die Beobachtung, daß ein Wort in einer Konstruktion in seiner üblichen Bedeutung verwendet wird, auf diesem Wort aber nicht die sonst für diese Konstruktion möglichen Transformationen ausgeführt werden können, sind Grundlage für die Tests. Dabei werden auf potentiellen Komposita eine Reihe von Transformationen ausgeführt und die potentiellen Komposita aufgrund der Akzeptabilität dieser Transformationen geordnet und klassifiziert. G. Gross[3] arbeitet mit ähnlichen Tests und stellt fest, daß der Grad der Lexikalisierung eines potentiellen Kompositums umgekehrt proportional ist zu der Anzahl der beobachtbaren transformationellen Eigenschaften:

> Il y a une relation de proportionnalité inverse entre le figement d'un groupe et le nombre de propriétés transformationnelles observables[4].

Obwohl auch dieses Kriterium nicht auf ungeteilte Zustimmung stößt[5], findet es u. a. in dem am LADL[6] erstellten Lexikon, einem der größten Wörterbuchprojekte in Frankreich, Anwendung (siehe unter Abschnitt 5.4.2 Das INTEX-System, Seite 85).

5.1.1.2 Status

Der Status eines Kompositums bezieht sich auf seine Verwendung, d. h. seine Frequenz, seinen Bekanntheitsgrad und den Bereich, in dem es verwendet wird. Der Status als Definitionskriterium ist nicht objektiv zu erfassen und außerdem Schwankungen in Raum und Zeit unterworfen. Habert und Jacquemin geben Beispiele für Komposita, die in einem bestimmten Zeitraum häufig verwendet wurden und allgemein bekannt waren bzw. in verschiedenen Verwendungskontexten nicht denselben Status und dieselbe Bedeutung haben.

---

1. Habert & Jacquemin 1993, S. 7.
2. U. a. Gross M. 1988.
3. U. a. bei Gross G. 1988.
4. Gross G. 1988, S. 23.
5. Vgl. dazu Habert & Jacquemin, 1993.
6. Laboratoire d'Automatique Documentaire et Linguistique, Paris.

So verstand man beispielsweise in der Nachkriegszeit unter der als *personnes déplacées* (dt.: Vertriebene) bezeichneten Personengruppe eine Gruppe von Menschen mit speziellem administrativem Status, was sich auch in der Tatsache zeigt, daß eine Abkürzung dafür existiert (*DP*, laut Habert und Jacquemin). Es ist unklar, ob ein Leser in der heutigen Zeit die Sequenz *personnes déplacées* als Beschreibung oder als Bezeichnung auffaßt.

Die Beurteilung des Status einer Sequenz ist auch vom Kontext bzw. der Gesprächssituation abhängig. So bezeichnet *pied de cuve* in einem allgemeinsprachlichen Kontext den Fuß eines großen Behälters (Wanne, Bottich), im Kontext der Weinherstellung ist es aber ein Fachausdruck für einen speziellen Gärbehälter.

Habert und Jacquemin erwähnen mehrere Tests[1], mit denen überprüft werden kann, ob ein Sprecher eine bestimmte Sequenz als eine kodierte Bezeichnung auffaßt. Als weitere Kriterien für kodierte Bezeichnungen gelten die Existenz von Abkürzungen, die Verwendung der Sequenzen in Indizes, Thesauri und Inhaltsverzeichnissen und der Gebrauch von Anführungszeichen. Trotz all dieser Kriterien bleibt die Beurteilung des Status einer Sequenz fast immer problematisch:

> La pratique est souvent d'appeler "noms composés" en langue générale ou "termes" en langue spécialisée des séquences stabilisées de manière assez nette dans une collectivité langagière donnée. Mais comme il s'agit d'un jugement sur l'usage, on comprend tout de suite son incontournable fragilité[2].

5.1.1.3 Bildungsverfahren

Die Untersuchungen der Bildungsverfahren von Komposita sind in einem Bereich zwischen Morphologie und Syntax angesiedelt. Einerseits müssen die Bildungsverfahren der Nominalkomposita in Zusammenhang mit anderen Wortbildungsverfahren (Präfigierung, Affigierung und Konversion) definiert werden, andererseits in Beziehung zur Generierung komplexer Sequenzen durch die Syntax gesetzt werden.

Hinsichtlich der Stellung der Wortbildung zwischen Morphologie und Syntax gibt es sehr unterschiedliche Positionen.

- **Trennung zwischen Syntax und Wortbildungsregeln**: Nominalkomposita werden nach anderen Regeln gebildet als Nominalphrasen. Bei der Bildung der Nominalkomposita kommen spezielle Verkettungsregeln zur Anwendung, die ein Beweis für eine Ebene sind, die unabhängig von der Syntaxebene ist, die im Rahmen des Satzes definiert wird.

  > Les syntagmes nominaux sont des *constituants* unis par des rapports d'*ordre syntaxique*, les composés nominaux sont des *formants* assemblés par des liens d'*ordre compositionnel*[3].

1. Habert & Jacquemin 1993, S. 10.ff
2. Habert & Jacquemin 1993, S. 10.

- **Enger Zusammenhang zwischen Syntax und Wortbildung**: Jede aus Lexemen gebildete Struktur ist das Produkt von Syntaxregeln, unabhängig davon, welcher lexikalische Status dem entstandenen Produkt zugewiesen wird. Die Bestimmung dieses Status ist vom Interpretationsmodus abhängig, der auf eine gegebene Struktur angewandt wird. Die Interpretation kann unter dem Aspekt der Kompositionalität (*compositionnalité*) oder unter dem Aspekt der Idiomatizität (*idiomaticité*) erfolgen.
- **Komposita als elliptische Sätze**: Diese Hypothese hat ihre Wurzeln in den Ausführungen von Darmesteter (1893) ("Un mot composé est une proposition en raccourci") und von Benveniste (1974) ("la composition nominale est une microsyntaxe"). In die gleiche Richtung geht die Hypothese der Transformationsgrammatik, nach der versucht wird, eine Verbindung zwischen Komposita und den Basissätzen herzustellen, aus denen die Komposita durch Transformationen hervorgegangen sein sollen. Die Zahl und die Komplexität der postulierten Transformationen läßt, von wenigen Einzelfällen abgesehen, diese Betrachtungsweise aber nicht sehr plausibel erscheinen.

5.1.1.4 Semantische Analyse: Transparenz vs Undurchsichtigkeit

In der Diskussion über die Semantik der Komposita trifft man auf zwei Extrempositionen. Entweder gelten die Komposita als opak, d. h. ihre Bedeutung kann aus den Bedeutungen ihrer Bestandteile nicht bestimmt werden, oder die Semantik der Komposita gilt als kompositionell, d. h. die globale Bedeutung eines Kompositums kann aus den Bedeutungen der konstituierenden Elemente konstruiert werden. Wie für die anderen Kriterien so gilt auch hier, daß die meisten Arbeiten die Komposita auf einem Kontinuum anordnen, daß von der totalen Undurchsichtigkeit bis zur offensichtlichsten Transparenz verläuft.

- In den Arbeiten des LADL wird neben der Bezeichnung "feststehender Ausdruck" (frz. *expression figée*), quasi als Synonym, auch die Bezeichnung "semantisch nicht zerlegbarer Ausdruck" (frz. *expression sémantiquement non compositionnelle*) verwendet:

  > Le sens des mots n'intervient pas dans l'interprétation des expression figées, elles sont donc apprises par coeur.[1]

  Diese Haltung erklärt sich daraus, daß die Semantik nach der Auffassung des LADL ein Bereich ist, der sich nicht formalisieren läßt und deshalb keine reproduzierbaren Beurteilungen von Phänomenen zuläßt. Die semantische Undurchsichtigkeit wird zwar postuliert, aber nicht weiter präzisiert.
- In den Arbeiten von Downing (1977) wird klar unterschieden zwischen historischen Phänomenen - d. h. die zahlreichen Abweichungen und Idiosynchrasien, die durch den Verlust der ursprünglichen Motivation oder durch neue, z. T. wider-

3. Habert & Jacquemin 1993, S. 12.

1. M. Gross 1988.

sprüchliche Motivation einer Sequenz entstanden sind - und den besonderen Interpretationsmöglichkeiten, die Komposita durch ihre Bildung haben. Von Einzelbeispielen schwierig zu interpretierender Komposita wird oft auf die allgemeine semantische Undurchsichtigkeit von Komposita geschlossen. Downing formuliert das Problem folgendermaßen:

> Un composé peut être fortement transparent sémantiquement quand il est forgé, mais une fois accepté par la communauté comme un nom conventionnel, il peut devenir aussi arbitraire que n'importe quel mot monomorphémique[1].

5.1.2 Fazit

Der Überblick über die Kompositaforschung konnte nur die Punkte darstellen, um die sich die Diskussion bewegt. Er sollte

- die Vielschichtigkeit der Problematik hinsichtlich der in den einzelnen Ansätzen angeführten Argumente und Kriterien aufzeigen,
- verdeutlichen, daß über keines der genannten Kriterien Einigkeit herrscht und daß
- die meisten Kriterien auf einem Kontinuum mit graduellen Differenzierungen angesiedelt werden müssen.

5.2 Computerlinguistische Arbeiten

5.2.1 Ansatz von Sta

Die Arbeiten von Sta[2]gehen folgender Frage nach: Wie kann aus einem bestehenden Korpus das Fachvokabular automatisch erstellt werden? Es geht ihm nicht darum, den Terminologieexperten zu ersetzen, sondern ihn möglichst gut bei seiner Aufgabe durch die Bereitstellung potentieller Fachtermini zu unterstützen.

Methodisch unterscheidet Sta den linguistisch und den statistisch basierten Ansatz, die, je nach Situation, zusammenarbeiten, nebeneinander bestehen oder sich ausschließen. Sta zeigt in seiner Studie eine mögliche Synthese zwischen beiden Ansätzen. Er wendet verschiedene Techniken aus der Statistik auf lexikalische Einheiten an, die durch eine linguistische Analyse aus einem Korpus extrahiert worden sind. Die statistischen Verfahren sollen Kriterien für die Bewertung der lexikalischen Einheiten liefern.

Laut Sta ist der Begriff "Fachausdruck" bzw. "Fachterminus" ursprünglich ein Konzept aus der Dokumentationswissenschaft, er wird aber immer mehr auch ein Untersu-

1. Downing 1977, S. 820.
2. Sta 1995.

chungsgegenstand in der Linguistik unter der Bezeichnung Fachterminus (*terme*), komplexe Benennung (*dénomination complexe*) oder Kompositum (*mot composé*).

Die linguistischen Ansätze beschreiben in erster Linie die formalen Eigenschaften der Fachtermini. Es gilt die Regel, daß der Fachterminus die Struktur einer Nominalgruppe hat. Aber diese Regel hat Ausnahmen (z. B. die Struktur "Nomen Nomen Nomen": *interface homme machine*) und gilt nicht in allen Bereichen (z. B. nicht für die Symbole in der Chemie). Der Autor hat einen aus 20 000 Fachtermini bestehenden Thesaurus[1] untersucht und die Strukturen von Nominalgruppen und deren Häufigkeit festgestellt, wobei Strukturen mit einer Häufigkeit von weniger als 2 % nicht berücksichtigt wurden. Dabei kommt er zu folgendem Ergebnis[2]:

Tabelle 16. Syntaktische Strukturen im Thesaurus der EDF

Syntaktische Struktur	Beispiel	%
Nomen Adjektiv	érosion fluviale	25.1
Nomen Präposition Nomen	analyse de contenu	24.4
Nomen	décentralisation	18.1
Eigennamen	Chinon	6.8
Nomen Präposition Determinator Nomen	assurance de la qualité	3.2
Nomen Präposition Nomen Adjektiv	unité de bande magnétique	2.8
Nomen Partizip Perfekt	puissance absorbée	2.2
Nomen Nomen	accès mémoire	2.1

Sta extrahiert aus einem Korpus des gleichen Gebiets Nominalgruppen, die diesen 7 Bildungsmustern entsprechen: sie werden als **potentielle Fachtermini** des Korpus (*candidats termes*) bezeichnet und stellen eine Untermenge der Fachtermini des Textes dar. Ihre Evaluierung erfolgt durch einen Abgleich mit einem Thesaurus desselben Fachgebiets. Das Korpus besteht aus 2000 technischen Texten, die jeweils eine Länge von 1-2 Seiten haben (830 000 Wörter insgesamt).

Über den Thesaurus werden zwei Hypothesen aufgestellt: er ist komplett und konsistent, d.h. ein potentieller Fachterminus, der im Thesaurus enthalten ist, gilt als korrekt, ein potentieller Fachterminus, der nicht im Thesaurus enthalten ist, gilt als falsch. Die Intersektion zwischen den extrahierten potentiellen Termini und dem Thesaurus

1. Thesaurus der EDF (Electricité de France).
2. Zitiert nach: Sta 1995, S. 121.

ergibt ein Verhältnis von 1:10, d. h. nur 10 % der nach rein linguistischen Kriterien extrahierten potentiellen Fachtermini sind im Thesaurus enthalten.

Die Tatsache, daß nur 10 % der potentiellen Fachtermini im Thesaurus enthalten sind, rechtfertigt nicht nur die Anwendung weiterer Filter, sie erfordert sie sogar. Sta verwendet vier statistische Filter: die Frequenz, die Varianz, den Diskriminationswert und die lokale Dichte. Die Filter berechnen für jeden potentiellen Terminus einen statistischen Wert. Eine Filtermethode ist umso effizienter, je mehr sie es ermöglicht, aus der Menge der potentiellen Termini diejenigen herauszufiltern, die im Thesaurus enthalten sind.

Insgesamt sind die Varianz und die lokale Dichte die besten Filter, die Frequenz liegt an dritter Stelle. Allerdings fällt diese Bewertung je nach Textlänge unterschiedlich aus, manche Filter eignen sich mehr für kürzere, andere wiederum mehr für längere Texte.

5.2.2 Ansatz von Bourigault

LEXTER (Logiciel d'Extraction de Terminologie) ist ein Software-Paket für die automatische Erstellung von Terminologie aus französischen Fachtexten[1]. LEXTER extrahiert potentielle terminologische Einheiten[2], die einem Experten zur Validierung vorgelegt werden. Die Vollständigkeitsquote des Systems soll so hoch wie möglich sein, da es leichter ist, einen potentiellen Terminus bei der Evaluierung zu eliminieren als Fachtermini zu finden, die überhaupt nicht extrahiert wurden.

Die Extraktion erfolgt in zwei Schritten: Analyse (Splitting) und Parsing. LEXTER wendet eine Regelbasis an, die Grenzen zwischen Nominalphrasen lokalisiert. In einem zweiten Schritt werden diese Nominalphrasen geparst, um daraus Teilphrasen zu extrahieren, die aufgrund ihrer grammatikalischen Struktur mögliche terminologische Einheiten darstellen.

Die semantische Funktion eines Fachterminus liegt in der Repräsentation eines Konzepts eines Fachgebiets. Bourigault schreibt dem Terminus Kontextunabhängigkeit zu, d. h. er sieht eine 1:1-Entsprechung zwischen einem linguistischen Ausdruck und einem extralinguistischen Objekt. Jeder Terminus hat seinen Platz in einem Netzwerk von Termini, das die Konzepte des Fachgebiets darstellt. Diese Referenzfunktion ist nach Benveniste (1974) die *synaptische* Markierung eines Syntagmas. Die Funktion des Terminus, ein Konzept außerhalb jedes Kontexts eindeutig darzustellen, hat nach Bourigault zur Folge, daß die Bildung von Termini einer Reihe von Beschränkungen unterliegt. Diese syntaktischen Bildungsmuster werden auch *synaptische Komposition*

1. Bourigault 1994.

2. Drozd & Seibicke 1973 (S.146) verwenden die Bezeichnung "terminologische Einheiten" als Oberbegriff für alle Formen von Fachtermini, z. B. Wortzusammensetzungen, Wortableitungen, Kürzungen, komplexe Phrasen usw.

genannt (Benveniste 1974). Beispielsweise sind französische Fachtermini im allgemeinen aus Nomen und Adjektiven gebildet, enthalten praktisch nie konjugierte Verben, die gebräuchlichsten Präpositionen sind “de” und “à”, auf die nur selten ein Determinator folgt.

Aus diesen Betrachtungen über die Form und die Funktion von Fachtermini folgert Bourigault:

- Ein Terminologie-Extraktionsprogramm kann allein auf syntaktischen Daten basieren, da die grammatische Form von Fachtermini relativ vorhersehbar ist.
- Es ist nicht zu erwarten, daß ein solches Programm **nur** Fachtermini extrahiert. Die von einem solchen Programm ausgegebenen Einheiten sind als **potentielle** Fachtermini zu betrachten.

Beim Splitting (erste Phase) nützt LEXTER “negatives” Wissen über die Form von Fachtermini, indem grammatikalischen Muster lokalisiert werden, die nie in einem Fachterminus enthalten sind und so mögliche Begrenzer von Fachtermini sind. Solche Muster sind z. B. konjugierte Verben, Pronomen, Konjunktionen, Verbindungen zwischen Präposition und Determinatoren etc. Das Splittingmodul enthält eine Regelbasis für die Lokalisierung von Begrenzern und hat als Eingabe ein lemmatisiertes und mit Wortklassen annotiertes Korpus. Seine Ausgabe besteht in einer Menge von Textsequenzen, von denen die meisten Nominalphrasen sind. Diese Nominalphrasen sind entweder schon potentielle Fachtermini (z. B: *traitement de texte*) oder sie enthalten Teilphrasen, die potentielle Fachtermini sind (z. B.: *disque dur de la station de travail* enthält *disque dur* und *station de travail*). Die vom Splittingmodul ausgegebenen Sequenzen werden Nominalphrasen maximaler Länge bzw. maximale Nominalphrasen genannt (*Maximal Length Noun Phrases: MLNP*).

Beim Parsing (zweite Phase) werden die maximalen Nominalphrasen mit Hilfe einer Regelbasis in kleinere Teilphrasen zerlegt und nach Head (H) und Modifier (M) analysiert. Das Parsingmodul erkennt ca. 800 verschiedene Strukturen und kann so fast 95% der maximalen Nominalphrasen des Testkorpus (1.2 Mio Wörter) analysieren.

Ein Beispiel für eine Regel[1]

N1 Adj Prep Det N2 Prep N3

-> N1 Adj

-> N2 Prep N3

z. B: disque dur de la station de travail

-> disque dur

H: disque

M: dur

-> station de travail

H: station

M: travail

Für die korrekte Lösung einiger problematischer Splitting- und Parsingfälle benötigt das System syntaktische Information zur Subkategorisierung. Bourigault[2] gibt dazu das folgende Beispiel:

(1) une armoire de contrôle *sensible* / **à une** / élévation de température

(1') une (armoire de contrôle) (sensible à une élévation de température)

Das Splittingmodul würde normalerweise die Sequenz *à une* als Begrenzer interpretieren und eine falsche Zerlegung wie in (1) vornehmen.

Für eine richtige Zerlegung benötigt das System in diesem Fall zusätzliche syntaktische Informationen über die Subkategorisierung des Adjektivs *sensible*. Derartige Informationen stehen dem System nicht a priori zur Verfügung. Das System ist jedoch mit einer Lernprozedur ausgestattet (*corpus-based endogeneous learning procedure*), die es ihm ermöglicht, sich solche Subkategorisierungsinformationen aus dem Korpus "selbst" anzueignen. Für die Lösung des o. g. Falls werden in einem ersten Durchgang alle Adjektive gesammelt, die in prädikativer Stellung stehen und auf die die Präposition *à* folgt. In einem zweiten Durchgang wird jedes Mal, wenn das Splittingmodul eine mit der Präposition *à* beginnende Sequenz eliminiert hat, das vorangehende

1. Bedeutung der Kategorien: N1, N2, N3: Nomen; Prep: Präposition; Det: Determinator; Adj: Adjektiv.

2. Bourigault 1996.

Adjektiv eliminiert, wenn es zu der erstellten Liste gehört. Für die Subkategorisierung von Nomen, Adjektiven und Partizip-Perfekt-Formen im Hinblick auf die gängigsten Präpositionen wurden Lernprozeduren implementiert, die i. a. komplexer sind als das angeführte Beispiel.

Das Strukturierungsmodul (dritte Phase) nutzt die vom Parsingmodul bereitgestellten Informationen zur Anordnung der extrahierten potentiellen Fachtermini in einem Netzwerkformat, das **terminologisches Netzwerk** (*réseau terminologique*) genannt wird. Dieses Modul stellt für jeden analysierten komplexen Terminus eine Verbindung zu den potentiellen Termini her, die denselben Head oder denselben Modifier enthalten. Dieses Hypertextnetz erleichtert dem Terminogieexperten die Validierung der Termini.

Evaluierung

LEXTER wurde mit verschiedenen Korpora getestet und evaluiert[1]. Das Splitting- und Parsingmodul erkennen und analysieren in der Regel 90-95 % der Nominalphrasen eines Textes korrekt. Die Analyseergebnisse aus einem Teilkorpus (20 000 maximale Nominalphrasen) wurden unter dem Aspekt der Art und der Frequenz der darin vorkommenden Nominalphrasen untersucht.

Unter den 20 000 extrahierten maximalen Nominalphrasen sind ca. 10 000 verschiedene Nominalphrasen, die nach 1875 verschiedenen Bildungsmustern konstruiert sind. Die Verteilung der Nominalphrasen auf die verschiedenen Bildungsmuster ist sehr unterschiedlich[2]:

- Die drei häufigsten Bildungsmuster stellen ein Viertel der zu analysierenden Nominalgruppen.
- Die 15 häufigsten Bildungsmuster decken 50 % des Textes ab.

1. Bourigault 1994, Kap. V.
2. Bourigault 1994, S.182 ff.

Tabelle 17. Grammatische Kategorien zur Beschreibung der Bildungsmuster

Kategorie	Erklärung
nom	Nomen und unbekannte Wörter
adj	Adjektive, sowie Partizip Perfekt und Partizip Präsens
de	die Präposition “de”
à	die Präposition “à”
prep	die Präpositionen “avec”, “contre”, “dans”, “par”, “pour”, “sous”, “sur”, “vers”
le	der bestimmte Artikel (“le”, “la”, “les”, “l’”) bzw. die entsprechenden elidierten Formen

Tabelle 18. Die 20 produktivsten Bildungsmuster maximaler Nominalphrasen

Rang	Frequenz	Struktur
1	1243	nom adj
2	903	nom de nom
3	716	nom de le nom
4	337	nom nom
5	300	nom de nom adj
6	228	nom de le nom adj
7	209	nom prep nom
8	201	adj nom
9	199	nom de le nom de nom
10	158	nom adj adj
11	138	nom adj de le nom
12	135	nom de nom de le nom
13	127	nom adj de nom
14	101	nom de nom de nom
15	80	nom de le nom nom
16	73	nom de le nom de le nom
17	72	nom à le nom
18	61	nom adj nom
19	60	nom de nom nom
20	52	nom prep de nom

5.2.3 Ansatz von Daille

Die Arbeiten von Béatrice Daille[1] wurden im Rahmen eines Projekts realisiert, in dem automatische Verfahren für die Erstellung von monolingualen bzw. bilingualen terminologischen Datenbanken entwickelt wurden. Das beschriebene Verfahren kombiniert linguistische Methoden mit statistischen Filtern. Das untersuchte Korpus stammt aus dem Bereich der Telekommunikation, die extrahierten Termini wurden durch einen Abgleich mit einer Terminologiedatenbank aus demselben Bereich evaluiert.

Fachvokabular setzt sich laut Daille aus einfachen und komplexen lexikalischen Einheiten zusammen. Komposita machen den größten Teil dieser komplexen lexikalischen Einheiten aus. Fachtermini in der Form einfacher Nomen werden in Dailles Arbeit nicht berücksichtigt.

Das Verfahren gliedert sich in zwei Arbeitsschritte:

- Extraktion potentieller Fachtermini (definiert durch morphosyntaktische Bildungsmuster)
- Bewertung der potentiellen Fachtermini mit Modellen aus der Statistik

Bei der Definition der potentiellen Fachtermini geht Daille empirisch vor. Aus einem satzweise alignierten zweisprachigen (engl.-frz.) Fachtext zur Telekommunikation wurden manuell einige hundert Fachtermini extrahiert und nach linguistischen Kriterien analysiert und klassifiziert. Die daraus resultierende linguistische Spezifikation ist die Grundlage des Extraktionsprogramms. Nach Daille sind Fachtermini in der Regel komplexe lexikalische Einheiten nominalen Typs und können als eine **Unterklasse der Komposita** betrachtet werden. Sie werden nach ihrer morphosyntaktischen Struktur klassifiziert und können so in die von M. Mathieu-Colas (1988) ausgearbeitete allgemeine Typologie der Komposita eingeordnet werden. Daille nennt drei Kriterien für Fachtermini: sie haben die morphosyntaktische Form eines Kompositums, gehören zum technischen Vokabular und werden immer gleich übersetzt.

> Seront considéres comme termes de notre domaine les séquences: épousant une structure morphosyntaxique d'un nom composé, intégrant un vocabulaire technique et possédant une traduction unique.[2]

Die Fachtermini werden nach ihrer Länge und ihrem Bildungsmuster (Abfolge von Wortklassen) beschrieben. Fachtermini der Länge 1 sind im wesentlichen Bindestrichwörter, z. B.: *semi-conducteur*, *court-circuit*, *plate-forme*. Fachtermini der Länge 2 sind bei weitem die häufigsten. Sie gelten als Basistermini (*termes de base*), aus denen die drei- und mehrgliedrigen Termini gebildet werden. Die zweigliedrigen Termini werden nach folgenden Mustern[3] gebildet:

1. Daille 1994.
2. Daille 1995, S. 104.
3. N, N1, N2: Nomen; Adj: Adjektiv; Det: Determinator; Prep: Präposition.

N Adj: *station brouilleuse*

N1 N2: *diode tunnel, station service*

N1 *à* (Det) N2: *antenne à réflecteur, assignation à la demande*

N1 *de* (Det) N2: *modulation de fréquence, synchronisation des paquets*

N1 Prep[1] N2: *multiplexage en fréquence*

Drei- und mehrgliedrige Termini sind selten und wurden nicht extrahiert.

Das Korpus stammt aus dem Bereich der Telekommunikation und umfaßt ca. 1 Mio. Wörter. Es wurde lemmatisiert und mit Wortklassen annotiert. Ein Extraktionsprogramm lokalisiert die morphosyntaktischen Sequenzen im Text, die den Bildungsmustern potentieller Fachtermini entsprechen. Die zur Extraktion verwendeten Automaten erkennen die beschriebenen Muster der Länge 2. Die lokalisierten Sequenzen sind die Eingabe für das zweite Modul: Anwendung statistischer Filter.

Alle extrahierten Sequenzen sind Komposita, aber nicht jedes Kompositum ist ein Fachterminus. Die statistische Analyse bewertet die extrahierten Sequenzen und soll eine Antwort auf die Frage liefern, welche Sequenz möglicherweise ein Fachterminus ist. Für jede Sequenz werden außer der Frequenz eine Reihe von Assoziationswerten und Distanzmaßen berechnet, die die Kohäsion der Elemente der Sequenz evaluieren: *Mutual Information*, Diversität nach Shannon etc. Eine extrahierte Sequenz gilt dann als Fachterminus, wenn sie in einer terminologischen Datenbank des entsprechenden Fachgebiets enthalten ist.

Die Frequenz erweist sich als sehr guter Indikator für den terminologischen Charakter eines Kompositums, ist aber problematisch bei selteneren Komposita. Die Grenze, ab der die Frequenz ein sicheres Kriterium ist, liegt sehr hoch. *Mutual Information* erwies sich wider Erwarten nicht als eine brauchbare Filtermethode. Der Ähnlichkeitsquotient nach Dunning (1993) zeigte sehr gute Ergebnisse.

5.3 Studie 1: Linguistische Beschreibung potentieller Fachtermini

Ein Ausschnitt aus dem EPÜ wurde als Grundlage für eine linguistische Beschreibung potentieller Fachtermini ausgewählt. Dieser Text hatte in der Alignierung hervorragende Ergebnisse gezeigt (siehe unter Abschnitt 4.4.2 Testreihen und Ergebnisse, Seite 53). Im EPÜ werden die Grundlagen der Europäischen Patentorganisation präzise und “verbindlich” beschrieben. Zu dieser Darstellung gehört die Einführung und Definition verschiedenster juristischer, administrativer und technischer Begriffe. Die Struktur der einzelnen Artikel ist meist dergestalt, daß im Titel des Artikels ein neuer Begriff eingeführt wird, der dann im darauf folgenden Text definiert und erläutert

1. Gemeint sind Präpositionen außer “de” und “à”, die gesondert behandelt werden.

wird. Ein großer Teil dieser Begriffe sind potentielle Fachtermini. Die eingeführten Begriffe sind teilweise sogar explizit als solche gekennzeichnet:

> Les brevets délivrés en vertu de la présente convention sont *dénommés* brevets européens[1].
>
> Il est institué par la présente convention une Organisation européenne des brevets, ci-après *dénommée* l'Organisation[2].

Aus diesen Gründen ist dieser Ausschnitt aus dem EPÜ besonders für eine Studie geeignet, in der die linguistischen Charakteristika der potentiellen Fachtermini des Gebiets herausgearbeitet werden sollen.

5.3.1 Nominalphrasen maximaler Länge

Der Text wurde intellektuell geparst, d. h. die einzelnen Sätze und Überschriften wurden in Nominal- und Verbalphrasen zerlegt. Die Verbalphrasen wurden nicht weiter untersucht. Die lokalisierten Nominalphrasen sind **Nominalphrasen maximaler Länge**, vergleichbar mit den *MLNP* in Bourigault's Arbeit. Die Untersuchung dieser maximalen Nominalphrasen berücksichtigte folgende Aspekte:

- Welche Typen maximaler Nominalphrasen kommen vor?
- In welche Phrasen können die maximalen Nominalphrasen zerlegt werden?
- Welche Phrasentypen kommen vor?
- Wie ist die Frequenzverteilung der Phrasen?
- Welche Phrasen sind Komposita bzw. potentielle Fachtermini des Gebiets?
- Welche Phänomene können bei einer geplanten maschinellen Untersuchung berücksichtigt werden?
- Welche Phänomene können voraussichtlich maschinell nicht behandelt werden?

Maximale Nominalphrasen, die nur aus einem Determinator und einem Nomen bestehen, sind in diesem Text sehr selten. Etwa drei Viertel der maximalen Nominalphrasen müssen weiter zerlegt werden. Eine Untersuchung der 60 Überschriften der einzelnen Artikel und der Kapitelnamen zeigte, daß dort schon fast alle Typen der im Korpus attestierten maximalen Nominalphrasen vorkommen. Am häufigsten sind maximale Nominalphrasen, die eine Präpositionalgruppe enthalten und maximale Nominalphrasen, die aus einem Nomen und einem Adjektiv bestehen. Die Präposition "de" ist bei weitem die häufigste. Es kommen auch zusammengesetzte Präpositionen vor, auf die später unter dem Aspekt der Lemmatisierung noch eingegangen wird.

1. EPÜ, Artikel 2(1).
2. EPÜ, Artikel 4(1).

Maximale Nominalphrasen mit Präpositionalgruppe

(1) la demande *de* brevet européen: die europäische Patentanmeldung

(2) l’adoption *d*'instructions administratives internes: der Erlass interner Verwaltungsvorschriften

(3) la Grande chambre *de* recours: die Große Beschwerdekammer

(4) l’office européen *des* brevets: das Europäisches Patentamt

Es gibt allein 12 maximale Nominalphrasen, die Zusammenstzungen mit *office européen des brevets* sind:

(5) les agences *de* l'Office Européen *des* brevets: die Dienststellen des Europäischen Patentamts

(6) la direction *de* l'office européen *des* brevets: die Leitung des Europäischen Patentamts

Nominalphrasen mit komplexen Präpositionen:

(7) du Traité de Coopération *en matière de* brevets du 19 juin 1970: des Vertrags über die internationale Zusammenarbeit auf dem Gebiet des Patentwesens vom 19. Juni 1970

(8) les brevets délivrés *en vertu de* la présente convention: die nach diesem Übereinkommen erteilten Patente

Maximale Nominalphrasen, die aus einem Nomen und einem Adjektiv bestehen

(9) la portée territoriale: die territoriale Wirkung

(10) le statut juridique: die Rechtsstellung

(11) la division juridique: die Rechtsabteilung

(12) l’avis technique: das technische Gutachten

(13) le règlement financier: die Finanzordnung

Des weiteren kommen vor:

Maximale Nominalphrasen mit Partizipialkonstruktion

(14) un brevet national *délivré* dans cet Etat: ein in diesem Staat erteiltes nationales Patent

(15) dans le délai *prévu* par le règlement d'exécution: innerhalb einer in der Ausführungsordnung vorgeschriebenen Frist

Maximale Nominalphrasen mit einer von einem Adjektiv eingeführten Nominalgruppe

(16) la loi *applicable au* contrat en cause: das Recht, das auf den betreffenden Vertrag anzuwenden ist

(17) un droit *commun aux* Etats contractants en matière de délivrance de brevets d'invention: ein den Vertragsstaaten gemeinsames Recht für die Erteilung von Erfindungspatenten

Maximale Nominalphrasen mit Koordination

(18) des biens immobiliers *et* mobiliers: bewegliches und unbewegliches Vermögen

(19) dans un but d'information *ou* de liaison: Informations- oder Verbindungszwecke

5.3.2 Klassifikation der maximalen Nominalphrasen

Die maximalen Nominalphrasen wurden in kleinere, syntaktisch und semantisch zusammenhängende Einheiten zerlegt, die **Phrasen** genannt werden. Sie sind linguistisch korrekte Nominalphrasen und *können* Fachtermini eines Gebiets sein. Die im Testkorpus lokalisierten Phrasen wurden unterschieden in:

- einfache Nomen, z. B. *brevet*: Patent.
- Komposita, die eindeutig identifiziert werden können, z. B. *langue officielle*: Amtssprache.
- **komplexe Nominalphrasen**: das sind Strukturen, die nicht eindeutig zerlegt werden können. Komplexe Nominalphrasen sind weder einfache Nomina noch eindeutig identifizierbare Komposita. Es ist unklar, ob und wie sie weiter zerlegt werden können; z. B: *Office européen des brevets*: Europäisches Patentamt.

Zur Bestimmung der Frequenz und zur Ermittlung von Regelmäßigkeiten wurden die Phrasen, d. h. in erster Linie Komposita und komplexe Nominalphrasen, nach Länge und Bildungsmuster klassifiziert, in Anlehnung an die Arbeiten von Sta (1995), Daille (1994), und Gaussier (1995). Das heißt aber nicht, daß ihr Status in jedem Fall eindeutig geklärt ist. Deshalb wurde diese Klasssifikation als *vorläufige Klassifikation* bezeichnet.

5.3.2.1 Bestimmung der Phrasenlänge

Die Länge einer Nominalphrase wird aus der Anzahl der Hauptwortarten[1] - damit sind hier Nomen, Adjektiv und Partizip gemeint - bestimmt:

- Länge 1 (eingliedrig): brevet, Vice-Président, court-circuit
- Länge 2 (zweigliedrig): Chambre de recours, procédure orale
- Länge 3 (dreigliedrig): Office Européen des brevets, Chambre de recours juridique
- Länge 4 (viergliedrig): Organisation Mondiale de la Propriété Industrielle.

1. “Hauptwortarten” als Übersetzung für das frz. *unité lexicologique pleine* (Nomen, Adjektiv, Partizip Verb).

5.3.2.2 Bestimmung der Bildungsmuster

Die Phrasenbildungsmuster geben die Art und die Reihenfolge der grammatischen Kategorien an, z. B.

- NA: Nomen Adjektiv: procédure orale
- NdeN: Nomen *de* Nomen: chambre de recours.

Eine vollständige Liste der im Korpus berücksichtigten Bildungsmuster findet sich unter Abschnitt 5.4.3 Verwendete Suchmuster, Seite 87.

5.3.3 Komplexe Nominalphrasen

Bei komplexen Nominalphrasen ist zu entscheiden, ob sie

- als zusammenhängende Einheit, d. h. als Kompositum, zu behandeln sind oder
- weiter zerlegt werden müssen.

Diese Frage stellte sich als eines der größten Probleme bei der Bestimmung und Extraktion der Fachterminologie des Testkorpus heraus und wird deshalb ausführlich behandelt. Zuerst werden einige Fälle aus dem Korpus dargestellt, deren Zerlegung unklar war, danach werden zwei Arbeiten vorgestellt, die sich ausführlich mit diesem Problem beschäftigen.

5.3.3.1 Das Problem der Zerlegung: Modifikation von Komposita oder Neubildungen von Komposita aus bereits bestehenden Komposita

Bei der Zerlegung der Nominalphrasen maximaler Länge in Teilphrasen ist oft unklar, ob eine Teilphrase als ein Kompositum, als zwei Komposita oder als Modifikation eines schon bestehenden Kompositums zu bewerten ist. Diese Frage ist umso schwieriger zu beantworten, wenn die zwei folgenden Aspekte berücksichtigt werden:

- Die frz. Teilphrasen werden ihren dt. Übersetzungsäquivalenten gegenübergestellt.
- Die Semantik der Teilphrasen wird durch einen fachlichen, d. h. *hier* juristischen oder technischen Kontext bestimmt.

Die Frage, ob die isolierten Teilphrasen terminologischen Charakter haben, wird vorerst zurückgestellt.

Beispiele für unklare Zerlegungen:

(1) Office européen des brevets: Europäisches Patentamt

Hier handelt es sich um den Namen einer europäischen Institution, die Teilphrase muß deshalb als *eine* Einheit interpretiert werden. Diese Analyse wird gestützt durch die Tatsache, daß einzelsprachliche Abkürzungen für diese Phrase existieren (frz. *OEB*, dt. *EPA*).

(2) Grande chambre de recours: Große Beschwerdekammer

Die so bezeichnete juristische Institution ist in Art. 22 des EPÜ in ihrer Zusammensetzung und ihren Aufgaben beschrieben und grenzt sich somit inhaltlich klar von der mit *chambre de recours* (Beschwerdekammer) bezeichneten Institution ab, die ebenfalls in verschiedenen Artikeln des EPÜ beschrieben ist. Die Tatsache daß die Adjektive *Grande* bzw. *Große* hier groß geschrieben werden, spricht für diese Interpretation.

(3) demande de brevet européen: europäische Patentanmeldung

Diese Struktur scheint aus *brevet européen* (europäisches Patent), dem Titel des Artikel 2 des EPÜ und *demande de brevet* (Patentanmeldung), einem der häufigsten Komposita im untersuchten Korpus, entstanden zu sein. Zu *brevet européen* ist im Korpus ein ganzes Paradigma attestiert, zu dem z. B. *brevet national*, *brevet américain* etc. gehören. Es liegt nahe, *demande de brevet européen* als eine Einheit zu interpretieren, auch wenn linguistisch nicht geklärt werden, ob diese Bildung aus den zwei erwähnten Einheiten entstanden ist oder ob es sich um eine Modifikation von *demande de brevet* handelt.

Aus der Literatur zur frz. Kompositaforschung habe ich zwei Arbeiten herausgeggriffen, die dieses Problem ansprechen. Die von den Autoren angeführten Beispiele werden durch Beispiele aus dem EPA-Korpus ergänzt.

5.3.3.2 Jacquemin: Bildung neuer Komposita

Diese Arbeit[1] untersucht die Bildung neuer Komposita aus bereits bestehenden Komposita (*surcomposition)* und unterscheidet dabei zwei Verfahren.

Bei der **Juxtaposition** (*juxtaposition*) bleibt die Struktur des Kompositums oder der Komposita erkennbar. So wird z. B. aus dem Kompositum *champ éléctrique* ein neues Kompositum *champ éléctrique statique* gebildet durch Juxtaposition des Adjektivs *statique* neben das Kompositum *champ éléctrique*. Die Struktur des Kompositums *champ éléctrique* wird dadurch nicht verändert.

So wären die Phrasen in Beispiel (2) und Beispiel (3) als Juxtaposition zu interpretieren. In (2) wird das Adjektiv *grande* vor das Kompositum *chambre de recours* und in (3) das Adjektiv *européen* hinter das Kompositum *demande de brevet* gestellt. Die Struktur der beiden Komposita *chambre de recours* und d*emande de brevet* verändert sich dadurch nicht.

Bei der **Verschmelzung** (*recouvrement*) kann die Struktur eines schon existierenden Kompositums verändert werden. So wird z. B. aus den Komposita *panneau de comptage* und *comptage éléctrique* das neue Kompositum *panneau de comptage éléctrique* gebildet. In diesem Fall verändert sich die Struktur der beiden Komposita nicht, da die Verschmelzung in der Mitte stattfindet.

1. Jacquemin 1991.

Bei der Bildung von *antenne parabolique de réception* aus *antenne parabolique* und *antenne de réception* findet die Verschmelzung über den Kopf des Kompositums statt, die Struktur von *antenne de réception* wird verändert.

Beispiel (3) könnte nach dieser Definition auch als Verschmelzung der Komposita *demande de brevet* und *brevet européen* interpretiert werden, die beide im Korpus attestiert sind.

Beispiel (1) müßte als Verschmelzung der Komposita *office des brevets* und *brevet européen* gedeutet werden. Es ist aber unklar, warum dann das entstehende Kompositum nicht *office de brevet européen*, quasi parallel zu (3) *(demande de brevet européen*) heißt. Beispiel (1) ist auch nicht als Juxtaposition zu erklären, da das Adjektiv *européen* in der Mitte des attestierten Kompositums *office des brevets* steht.

Kritik an der Position Jacquemins:

Die Juxtaposition schließt die Modifikation eines Kompositums durch ein Adjektiv oder eine Präpositionalgruppe ein. Nach Jacquemins Definition entsteht dabei **immer** ein neues Kompositum, ein Kriterium, das teilweise problematisch ist. Es ist keinesfalls eindeutig, daß z. B. eine Bildung wie *demande de brevet européen* unbedingt als Kompositum zu bewerten ist, es könnte ebenso *nur* als Modifikation eines schon bestehenden Kompositums angesehen werden.

Die Definition der beiden Operationen Juxtaposition und Verschmelzung stützt sich auf das Vorkommen bereits bekannter Komposita, das bedeutet aber, daß zuvor ein Lexikon der Termini des Fachgebiets erstellt werden muß oder ein solches schon existiert.

5.3.3.3 Betrachtungsweise von Daille

In der Arbeit von Daille, die auch kontrastive Aspekte (frz.-engl.) berücksichtigt, werden drei Bildungsverfahren unterschieden, wie aus Komposita der Länge 2 Strukturen der Länge >= 3 entstehen:

- Neubildung von Komposita aus schon bestehen Komposita (*surcomposition*)
- Modifikation elementarer Komposita (*modification*)
- Koordination (*coordination*)

Im folgenden wird gezeigt, wie die drei Beispiele aus dem EPÜ-Korpus (siehe Abschnitt 5.3.3.1 Das Problem der Zerlegung: Modifikation von Komposita oder Neubildungen von Komposita aus bereits bestehenden Komposita, Seite 77) nach dem Klassifikationschema von Daille zu bewerten sind. Sie werden durch weitere Beispiele unklarer Zerlegungen aus dem EPÜ-Korpus ergänzt. Die von Daille angeführten Beispiele werden zusammen mit ihrer Übersetzung (engl.) zitiert.

5.3.3.3.1 Neubildung von Komposita aus schon bestehenden Komposita

Daille unterscheidet hier zwei Verfahren, die Juxtaposition und die Substitution. Die Juxtaposition erhält die interne Struktur des Basisterminus, bei der Substitution wird diese Struktur verändert.

1a. Juxtaposition

Eine durch Juxtaposition entstandene Neubildung hat folgende Eigenschaften:

- Die Elemente des Basisterminus bleiben nebeneinander bestehen.
- Es kommt mindestens ein Basisterminus vor.
- Die Juxtaposition vollzieht sich mit Hilfe einer Präposition.
- Die Überlappungen im Inneren der entstandenen Struktur beziehen sich nicht auf Komposita elementaren Typs.

Beispiele:

- Juxtaposition eines elementaren Kompositums und eines einfachen Nomens

  N1 Prep1 [N2 Prep2 N3]

  modulation par [déplacement de phase]: phase key shifting

  Die Strukturen *modulation par déplacement* und *modulation de phase* sind keine Basistermini.

aus dem EPÜ-Korpus:

  N1 Prep1 N2 Prep2 N3

  dispositions du règlement d'éxécution: Ausführungsordnung

  Die Struktur *dispositions du règlemen*t könnte evtl. als Kompositum interpretiert werden, der Status von *dispositions d'éxécution* ist unklar. Die Tatsache, daß die im Frz. entstandene Struktur im Dt. nur mit *einem* Kompositum übersetzt wird, spricht dafür, *dispositions du règlement d'éxécution* als ein Kompositum zu interpretieren.

- Juxtaposition von zwei Komposita elementaren Typs

  [N1 Adj1] Prep1 [N2 Prep2 (Det) N3]

  [accès multiple] avec [assignation à la demande]

  Die Strukturen *accès avec assignation* und *accès à la demande* sind keine Basistermini.

aus dem EPÜ-Korpus:

[N1 Adj1] Prep1 Det [N2 Adj2]

[Organisation Mondiale] de la [Propriété Industrielle]: Weltorganisation für geistiges Eigentum[1]

Die Strukturen *Organisation de la Propriété* und *Organisation Industrielle* sind keine Basistermini.

1b. Substitution

In einem Basisterminus der Länge 2 wird eines der beiden Wörter, die den Hauptwortarten angehören, durch ein Kompositum substituiert. Der Kopf des Kompositums ist das ersetzte Wort. Die Substitution unterscheidet sich von der Juxtaposition dadurch, daß

- obligatorisch zwei Basistermini darin vorkommen müssen und
- die interne Struktur einer der beiden Termini verändert werden kann.

Beispiel:

N1 Prep1 N2 + N1 Prep2 N3 -> N1 Prep2 N3 Prep1 N2

réseau à satellites + réseau de transit -> réseau de transit à satellites: transit satellite network

aus dem EPÜ-Korpus:

N1 Prep1 N2 + N2 Adj -> N1 Prep2 N2 Adj

demande de brevet + brevet européen -> demande de brevet européen: europäische Patentanmeldung

5.3.3.3.2 Modifikation

Die modifizierenden Elemente, die aus einem Basisterminus drei- oder viergliedrige Termini erzeugen, erscheinen entweder innerhalb, vor oder hinter diesem Basisterminus. Die entstandenen Termini *können* Komposita sein.

Die **Einfügung von Modifikatoren** bezieht sich vor allem auf Adjektive und Adverbien.

1. Der engl. Name der Organisation ist “World Intellectual Property Organisation” (WIPO).

Beispiele:

N1 Prep (Det) N2 -> N1 **Adj** Prep (Det) N2

réseaux **mondiaux** de télécommunication: global telecommunication networks

bzw.

N Adj -> N **Adv** Adj

réseaux **entièrement** numériques: all digital networks

aus dem EPÜ-Korpus:

Office européen des brevets: Europäisches Patentamt

Die **Postposition von Modifikatoren** bezieht sich auf Adjektive und adverbiale Präpositionalgruppen.

Beispiele:

[N1 N2] **Adj**

[interfaces usager-réseau] **polyvalentes**: multi-purpose user network interface

[N1 Adj1] [**Prep Adj N**]

[câble sous-marin] [**à large bande**]: wideband submarine cable

aus dem EPÜ-Korpus:

demande de brevet européen: europäische Patentanmeldung

Diese Struktur wurde schon als Substitution analysiert; sie könnte aber auch als Modifikation des häufig attestierten Kompositums *demande de brevet* interpretiert werden

Die **Voranstellung** von Modifikatoren unterscheidet sich von der Einfügung und Postposition dadurch, daß dabei meist keine neuen Komposita entstehen. Die vorangestellten Adjektive bilden eine geschlossene Klasse, die häufigsten unter ihnen sind die Adjektive *divers, nombreux, tel, autre, principal* etc. Die schon erwähnte Struktur *Grande chambre de recours* ist ein seltenes Beispiel für die Voranstellung von Modifikatoren, bei der ein neues Kompositum entsteht.

5.3.3.3.3 Koordination

Die Koordination wird wesentlich seltener als die beiden anderen Verfahren angewandt. Daille unterscheidet mehrere Formen der Koordination, auf die hier nicht weiter eingegangen wird. In vielen Fällen kann nicht entschieden werden, ob die entstandene Struktur als Kompositum zu bewerten ist oder nicht.

Beispiel:

> N1 de N3 + N2 de N3 -> N1 et N2 de N3
>
> équipement d'émission et équipement de transmission -> équipement d'émission et de transmission: transmit and receive equipment

aus dem EPÜ-Korpus:

> budget modificatif ou additionnel: Berichtigungs- und Nachtragshaushaltsplan
>
> bilan de l'actif et du passif: Übersicht über das Vermögen und die Schulden
>
> préparation et adoption du budget: Entwurf und Feststellung des Haushaltsplans.

5.3.3.3.4 Kritik

Die beschriebene Klassifikation hat selektivere Kriterien als das Schema von Jacquemin, sie wird aber allein dadurch nicht operationeller. Alle problematischen Fälle aus dem EPÜ können damit beschrieben werden, auch wenn, wie gezeigt, manchmal nicht nur eine Analyse denkbar ist. Auch dieses Schema arbeitet implizit mit einer im voraus erfolgten Bestimmung der zwei- und mehrgliedrigen Komposita des Fachgebiets.

Oft ist es unmöglich, bei einer Teilphrase klar zwischen der Neubildung eines Kompositums und der Modifikation eines schon bestehenden Kompositums zu unterscheiden. Da solche Teilphrasen i. a. seltener vorkommen als die darin enthaltenen Basistermini ist Dailles Kriterium der gleichbleibenden Übersetzung auch nicht immer anwendbar.

5.3.4 Fazit

Die linguistische Untersuchung des EPÜ-Korpus hat gezeigt, daß Komposita und Fachtermini zwar prinzipiell einer bestimmten Anzahl von Bildungsmustern entsprechen, daß dabei aber verschiedene Transformationen vorkommen können.

Die automatische Erkennung und Extraktion von Komposita ist aus folgenden Gründen problematisch:

- Es ist unmöglich, zu entscheiden, ob eine morphosyntaktische Struktur ein Kompositum der Länge 2 ist, bevor die Komposita der Länge 3 bestimmt worden sind.
- Umgekehrt ist es unmöglich, über den Status einer Sequenz der Länge 3 zu entscheiden, bevor die Komposita der Länge 2 bestimmt worden sind.

Dieses Problem kann in einer maschinellen Analyse nicht gelöst werden. Eine Heuristik sollte aber die Vorteile und Möglichkeiten eines großen Korpus nutzen. Deshalb

werden zweigliedrige Strukturen und die häufigsten Arten dreigliedriger Strukturen aus dem Text extrahiert. Es ist zu sehen, ob im Einzelfall die Auftretenshäufigkeit einer mehrgliedrigen Phrase und der darin enthaltenen Teilphrasen Hinweise zum Status der mehrgliedrigen Phrase geben kann.

Die Bestimmung der potentiellen Fachtermini aus dem vorliegenden bilingualen und parallelisierten Korpus kann über zwei Methoden erfolgen. Die erste Methode entspricht in etwa der von Bourigault (1994) ausgeführten Vorgehensweise, d. h. Splitting des Textes in maximale Nominalphrasen durch das Begrenzerverfahren und Parsing der maximalen Nominalphrasen in Teilphrasen. Die zweite Methode ähnelt den von Sta (1995) und Daille (1994) vorgestellten Verfahren, d. h. Extraktion von Nominalphrasen nach bestimmten Bildungsmustern.

Für die vorliegende Arbeit wurde aus verschiedenen Gründen die zweite Methode gewählt:

- Maximale Nominalphrasen, die nicht weiter geparst werden, abstrahieren nicht vom Einzelfall und stellen keine Verbindungen zwischen lexikalisch (und inhaltlich) ähnlichen Nominalphrasen her. Sie könnten deshalb nicht statistisch ausgewertet werden. Nur wenn Parser für beide Sprachen zur Verfügung stehen, können die maximalen Nominalphrasen in größerem Rahmen weiter zerlegt werden. In diesem Fall nähern sich die Ergebnisse der beiden Methoden an.
- Nominalphrasen, die bestimmten Bildungsmustern, insbesondere den Bildungsmustern von Komposita entsprechen, können einfacher extrahiert werden und stehen sofort für eine statistische Auswertung zur Verfügung. Der Text muß dazu nicht in seiner Gesamtheit geparst werden. Die Extraktion kann sich auch auf bestimmte, z. B. besonders häufige Bildungsmuster beschränken.

5.4 Studie 2: Suche nach einem geeigneten Werkzeug

5.4.1 Ziel der Studie

In der vorhergehenden Studie wurden die Einheiten beschrieben, die aus dem frz. Korpus extrahiert werden sollen. Da die Größe des Korpus die Grenzen einer rein intellektuellen Analyse sprengt, muß die Untersuchung maschinell mit einem oder mehreren geeigneten Software-Werkzeugen erfolgen. Das Ziel dieser zweiten Studie bestand darin, ein geeignetes Werkzeug zu bestimmen, dieses mit einem Teil des Korpus zu testen und eventuell für die gegebene Aufgabe anzupassen.

Eine maschinelle Analyse eines Textes setzt eine **Lemmatisierung** des Textes voraus, d. h. der Text wird in Wörter segmentiert und jedes Wort wird mit seiner (seinen) grammatischen Kategorie(n) und seiner (seinen) Grundform(en) ausgezeichnet. Es gibt verschiedene Methoden der Lemmatisierung[1]. Die Zuordnung von Wörtern, Lem-

1. Vgl. dazu Maier-Meyer 1995, Kap. 3.1.2.

mata und Kategorien erfolgt in den meisten Anwendungen durch “Nachschlagen” in einem elektronischen Wörterbuch.

Danach sollen im lemmatisierten Text bestimmte **Strukturen lokalisiert** werden. Diese Strukturen sind in diesem Fall Abfolgen von grammatischen Kategorien, die Bildungsmuster von Nominalphrasen darstellen.

5.4.2 Das INTEX-System

Das INTEX-System ist ein Softwaresystem[1] für die maschinelle Analyse von frz. Texten auf der Grundlage von Wörterbüchern, die am LADL (Laboratoire d’Automatique Documentaire et Linguistique) unter der Leitung von Prof. Maurice Gross entwickelt wurden. Die integrierten Wörterbücher und Grammatiken sind in Form endlicher Automaten repräsentiert. Der Benutzer hat die Möglichkeit, eigene Wörterbücher und Grammatikbeschreibungen zu erstellen und im Rahmen von INTEX zu benutzen. Zu INTEX gehört ein spezieller Editor, mit dem Automaten und Transducer (Automaten mit Ausgabe[2]) erstellt werden können. In einem lemmatisierten Text können Strukturen verschiedenster Art lokalisiert werden, z. B. die Vorkommen aller Formen eines Lemmas, morphosyntaktische Muster, syntaktische Strukturen etc. INTEX kann auch als Grundlage für ein Rechtschreibkorrekturprogramm eingesetzt werden.

5.4.2.1 Die INTEX-Lexika

Die integrierten Wörterbücher, das DELA-System (Dictionnaire éléctronique du LADL), sind *ein* Produkt eines umfangreichen Forschungsprogramms zur automatischen Analyse natürlicher Sprache. Die vollständige Beschreibung des gesamten frz. Vokabulars unter orthographischen, phonologischen, morphologischen und syntaktischen Aspekten ist das Hauptziel der seit mehr als 20 Jahren durchgeführen Forschungstätigkeiten. Das DELA-System besteht aus mehreren Teillexika, von denen das DELAS (Lexikon der einfachen Formen) und das DELAC (Wörterbuch der komplexen Formen) für die vorliegende Anwendung relevant sind.

Eine **einfache Form** ist orthographisch definiert als eine Sequenz von Buchstaben, die durch zwei Separatoren (Leerzeichen, Bindestriche, Apostroph, Punktuationszeichen, Tabulatoren etc.) begrenzt ist. Das heißt, daß auch Teile von Bindestrich-Komposita, die isoliert kein sinnvolles Lexem darstellen und kontrahierte Formen, die durch Apostroph abgetrennt sind, im DELAS (bzw. DELAF-Lexikon der einfachen flektierten Formen) aufgeführt werden müssen. **Komplexe Formen** bestehen aus mehreren Einheiten und sind nach den von Silberztein angegebenen Kriterien definiert.

Die flektierten einfachen und komplexen Formen wurden automatisch durch ein Programm generiert und in gesonderten Lexika abgespeichert. Das DELAF-Lexikon

1. Beschreibung in Silberztein 1993.
2. Beschreibung in Silberztein 1993 und Revuz 1991.

(*Dictionnaire des mots simples fléchis*) enthält etwa 700 000 flektierte einfache Formen, die aus 80 000 unflektierten einfachen Formen generiert wurden. Jeder Eintrag enthält für eine Wortform die dazugehörige morphologische Information: die Grundform, die Wortklasse und Flexionsinformation.

Beispiel:

a,avoir.V:P3s

abaissa,abaisser.V:J3s

abandons,abandon.N1:mp

Die Form *a* ist eine Form des Verbs *avoir,* sie ist konjugiert in der 3. Person Singular Präsens (*V:P3s)*. Die Form *abaissa* ist eine Form des Verbs *abaisser,* sie ist konjugiert in der 3. Person Singular Passé simple (*V:J3s).* Die Form *abandons* ist ein Nomen der Klasse N1, Maskulinum Plural (*N1:mp*).

Das DELACF-Lexikon (*Dictionnaire des mots composés flechis*) enhält ca.150 000 flektierte komplexe Formen, von denen der weitaus größte Teil allgemeinsprachlich ist. Komposita, die charakteristisch für die Patentdokumentation sind, sind im DELAC nicht enthalten. Jeder Eintrag enthält für eine Wortform die Grundform, die Wortklasse und Flexionsinformation.

Beispiel:

à/tout/de/suite,à/tout/de/suite.ADV+PCDC

pommes/de/terre,pomme/de/terre.N+NDN+Conc:fp

Die Sequenz *à tout de suite* ist ein unveränderbares Adverb der Klasse PCDC und hat *à tout de suite* als Grundform. Die Sequenz *pommes de terre* ist der Plural von *pomme de terre*, das Kompositum ist ein Femininum der Klasse NDN.

Durch Kompression der Daten, in diesem Fall durch Transformation des Lexikons in die Form eines endlichen Automaten[1], konnte der Umfang der Lexika erheblich reduziert werden. Das DELAF, mit einer Dateigröße von 10 Megabyte, kann auf 1 Megabyte reduziert werden. Mit dieser komprimierten Form des DELAF können auf OS/2-Systemen ca. 100 000 Wörter pro Minute nachgeschlagen werden.

5.4.2.2 Extraktion von Strukturen

In einem auf der Grundlage der INTEX-Lexika lemmatisierten Text können Suchmuster unterschiedlicher Art lokalisiert werden. Das entsprechende Menu des Programms heißt *Locate Pattern*. Suchmuster (frz. *patrons*, engl. *patterns*) können sein:

- ein Wort oder eine Wortliste, z. B. alle Futurformen des Verbs *faire,* alle Vorkommen des Nomens *brevet*.

1. Genaue Beschreibung des Verfahrens in Revuz 1991.

- eine grammatische Kategorie, z. B. alle Verben eines Texts, alle Nomen im Plural (*N:p*).
- eine morphosyntaktische Struktur, dargestellt in Form eines regulären Ausdrucks oder eines Automaten, z. B. <Nomen> (de + d') <N>.

Das Menu *Locate Pattern* hat einen lemmatisierten Text und ein Suchmuster als Eingabe. Die Ausgabe erfolgt als:

- alphabetisch geordnete Konkordanz oder als

- Text, in dem die lokalisierten Suchmuster in einem anderen Zeichensatz repräsentiert sind oder als

- Index, der die Anfangsposition und die Länge der lokalisierten Sequenz angibt. Aus dem Index kann die Satznummer rekonstruiert werden.

5.4.3 Verwendete Suchmuster

Im folgenden werden die Phrasenbildungsmuster beschrieben, die im Korpus berücksichtigt wurden. Die Muster beschreiben Nominalphrasen als Abfolge von grammatischen Kategorien. Sie werden **NP-Muster** genannt und in Form regulärer Ausdrücke dargestellt.

Die grammatischen Kategorien sind in der INTEX-Notation[1] angegeben. Darstellungen, die von dieser Notation abweichen, sind besonders gekennzeichnet. Ein NP-Muster wird durch einen regulären Ausdruck beschrieben und durch zweisprachige Beispiele aus dem Testkorpus ergänzt. Für die eigentliche Extraktion wurden die regulären Ausdrücke in Form von Automaten dargestellt (siehe Abschnitt 5.5.1.1 Beispiel: Darstellung der Strukturen als Automaten mit Kontextbeschreibungen, Seite 96).

5.4.3.1 NP-Muster der Länge 1

N: <N>

brevet: Patent

irrecevabilité: Unzulässigkeit

BS: Nomen mit Bindestrich

sous-revendication:Unteranspruch

Vice-Président: Vizepräsident

1. <N>: Nomen, <A>: Adjektiv, <Prep>: Präposition, <ms>: Maskulinum Singular, <mp>: Maskulinum Plural, <fs>: Femininum Singular, <fp>:Femininum Plural, "+": ausschließliches Oder; vgl. dazu Silberztein 1993.

5.4.3.2 NP-Muster der Länge 2

NdeN: <N> (de+d'+du+de la+de l'+des) <N>
demande de brevet: Patentanmeldung
division d'opposition: Einspruchsabteilung
titulaire du brevet: Patentinhaber
état de la technique: Stand der Technik
motifs de l'opposition: Einspruchsgründe
exposé des motifs: schriftliche Begründung

NA: (<N:ms><A:ms> + <N:mp> <A:mp> + <N:fs> <A:fs> + <N:fp> <A:fp>)
brevet européen: europäisches Patent
utilisation revendiquée: beanspruchte Verwendung

NàN: <N> (à+ au+à la+à l'+aux) <N>
stabilisants à la lumière: Lichtstabilisatoren
parties à la procédure: Verfahrensbeteiligte

NPrepN: <N> <PREP[1]> <N>
protection par brevet: Patentschutz
revêtement en teflon: Tefloninnenauskleidung

NN: <N> <N>
membre juriste: rechtskundiges Mitglied
valeur limite: Grenzwert
cation lanthane: Lanthankation

5.4.3.3 NP-Muster der Länge 3

NdeNA: <N> (de +d'+du+de la+de l'+des) <N> <A>
groupe d'Etats contractants: Vertragsstaatengruppe
chambre de recours juridique: Juristische Beschwerdekammer
chambre de recours technique: Technische Beschwerdekammer

1. Präpositionen außer "de" und "à".

NdeNdeN:
<N> (de+d'+du+de la+de l'+des) <N> (de+d'+du+de la+de l'+des) <N>
zéolites de silicate de bore: Borsilikatzeolithe
décision de révocation du brevet: Widerrufsentscheidung
procédé de conversion d'hydrocarbures: Kohlenwasserstoffumwandlungsverfahren

NAA: <N> <A> <A>
acide monocarboxylique aliphatique: aliphatische Monocarbonsäure
silicates métalliques cristallins: kristalline Metallsilikate
matières premières naturelles: natürliche Ausgangsstoffe
composition aqueuse non oxydante: nichtoxidierendes wäßriges Gemisch
composition pulvérisante orale: Mundspray

NAdeN: <N> <A> (de+d'+du+de la+de l'+des) <N>
Office européen des brevets: Europäisches Patentamt
solution aqueuse d'hexaméthylènediamine: wäßrige Lösung von Hexamethylendiamin
sulphoacétate laurique de sodium: Natriumlaurylsulfoacetat

5.4.3.4 NP-Muster der Länge 4

NdeNdeNA:
<N> (de +d'+du+de la+de l'+des) <N> (de+d'+du+de la+de l'+des) <N>
homme du métier de compétence moyenne: Durchschnittsfachmann
question de droit d'importance fondamentale: Rechtsfrage von grundsätzlicher Bedeutung
date de dépôt du brevet européen: europäischer Anmeldetag
atome d'azote du cycle pipéridine: Piperidinstickstoffatom

5.4.4 Probleme und Lösungen: Eignung für die Aufgabe und Anpassung

Zu Testzwecken wurde ein Teilkorpus von ca. 8000 Wörtern (zwei Dokumente aus den frz. Entscheidungen der Beschwerdekammer) mit INTEX in mehreren Durchgängen analysiert. So konnte festgestellt werden, welche Aufgaben mit INTEX gelöst werden können, welche Aufgaben nicht mit INTEX gelöst werden können, und welche Möglichkeiten der Anpassung im System vorgesehen sind.

5.4.4.1 Ambiguität

Bei der Lemmatisierung ordnet INTEX einer Wortform alle Kategorien zu, die für diese Form im Lexikon verzeichnet sind. Daraus resultiert ein hoher Grad an Ambiguität. So sind nur 50-80 % der Nominalphrasen, die aufgrund eines einfachen regulären Ausdrucks ohne weitere Umgebungsdefinition lokalisiert wurden, korrekt.

Sucht man beispielsweise in einem lemmatisierten Text nach Phrasen, die einem regulären Ausdruck der Form (<N> <N>) entsprechen, so erhält man Strukturen wie

(1) son siège

vorkommend in: L'Organisation a *son siège* à Munich.

(2) Convention pour

vorkommend in: ... la *Convention pour* la protection de la propriété industrielle signée ...

(3) il est

vorkommend in: *Il est* institué par la présente convention une ...

(4) examinateur technicien

vorkommend in: Les divisions d'examen se composent d'un *examinateur technicien*.

Wie der Test zeigt, werden bei der Suche nach <N> <N> alle Sequenzen von Wörtern lokalisiert, die Nomen sein können, es aber in der vorgefundenen Sequenz nicht unbedingt sind. Da viele Wörter *auch* Nomen sein können, ist die Korrektheit der extrahierten Sequenzen in diesem Fall besonders niedrig.

Die Ambiguität in Beispiel (1) könnte gelöst werden, indem im Linkskontext der Struktur ein Determinator gefordert wird.

Die Ambiguität in Beispiel (2) könnte gelöst werden, wenn *pour* nicht die Kategorie Nomen hätte (wie z. B. in dem komplexen Ausdruck *le pour et le contre*).

Die Ambiguität in Beispiel (3) könnte gelöst werden, wenn im Linkskontext der Struktur ein Determinator gefordert wird oder wenn *il* nicht die Kategorie Nomen hätte; ein Fall, der übrigens äußert selten ist.

Beispiel (4) ist eine korrekte Nomen-Nomen-Verbindung.

Die Analyse von Wortformen ergibt i. a. für einen Teil der untersuchten Formen ambige Ergebnisse. Ambiguitäten existieren, wenn auch in unterschiedlicher Ausprägung, in allen Sprachen und sind unabhängig von einem speziellen Lemmatisierungsverfahren. Es gibt verschiedene Strategien für die Behandlung lexikalischer und morphosyntaktischer Ambiguitäten. Die in der Literatur beschriebenen Disambiguierungsstrategien kann man unter zwei Aspekten charakterisieren. Einerseits stehen linguistisch orientierte Verfahren (sie werden auch als regelbasierte Verfahren bezeichnet) im Gegensatz zu statistisch orientierten Verfahren, andererseits lassen sich

wortbezogene Verfahren von kontext- bzw. satzbezogenen Verfahren unterscheiden. Eine Beschreibung der verschiedenen Strategien findet sich bei Maier-Meyer[1].

In der vorliegenden Arbeit ging es vor allem darum, Disambiguierungsstrategien zu finden, die unter Berücksichtigung der Korpusgröße und der von INTEX bereitgestellten Möglichkeiten realisierbar waren. Für ein echtes statistisch orientiertes Verfahren ist das Korpus zu klein, Frequenzverteilungen der Wortklassen einer Wortform konnten aber z. T. genutzt werden.

5.4.4.1.1 Lösungsmöglichkeit 1: Präferenzlexikon

Einige frequente Wörter, wie z. B. *a* (von *avoir*), *la*, *si* usw. sind ambig, weil sie eine spezielle Bedeutung in einem bestimmten Gebiet haben*: si* und *la* sind Nomen, wenn sie in der Musik als Bezeichnung für Noten verwendet werden, *a* ist ein Nomen, wenn es den Buchstaben "a" bezeichnet. Solche Wörter können in ein gesondertes Lexikon (Präferenzlexikon) eingetragen werden, in dem ihnen nur eine Wortklasse, die häufigste, zugeordnet wird. Bei der Benutzung von INTEX wird dann diesem kleinen Lexikon Priorität über das DELAS-Lexikon gegeben.

Beispiel:

der Eintrag für die einfache Form *il:*

il,il.N:ms und il.PRO:3ms (DELAS-Lexikon)

il,il.PRO:3ms (Präferenzlexikon)

der Eintrag für die einfache Form *pour*:

pour,pour.N:ms und pour, pour.PREP (DELAS-Lexikon)

pour, pour.PREP (Präferenzlexikon)

5.4.4.1.2 Lösungsmöglichkeit 2: Restriktion des Kontexts

Nominalphrasen sind durch bestimmte Typen grammatischer Kategorien begrenzt, wie z. B. Verben, Relativpronomen, Interpunktion etc.[2] In einer Reihe von Tests wurde versucht, die Umgebung (Rechts- und Linkskontext) für jeden zu extrahierenden Phrasenbildungstyp **möglichst optimal** zu bestimmen. Wird der Kontext zu restriktiv formuliert, sind zwar 100 % der extrahierten Phrasen korrekt, aber ein nicht zu unterschätzender Prozentsatz der Phrasen wird gar nicht extrahiert, weil ihre Umgebung nicht der Kontextdefinition entspricht (niedrige *Vollständigkeitsquote* und hohe *Genauigkeitsquote*). Ist die Kontextdefinition zu allgemein oder fehlt sie sogar ganz, dann ist ein zu großer Teil der extrahierten Phrasen schon unter linguistischen Gesichtspunkten nicht korrekt (hohe *Vollständigkeitsquote* und niedrige *Genauigkeitsquote*). Es können aber auch mehrere Kontextdefinitionen erstellt werden mit unter-

1. Maier-Meyer 1995, Kap. 3.5.
2. Wie von Bourigault 1994 beschrieben.

schiedlichem Grad an Präzision, je nachdem, welcher Parameter des Ergebnisses (*Vollständigkeitsquote* oder *Genauigkeitsquote*) optimiert werden soll.

In die Kontextdefinition fließen **allgemeine Syntax-und Grammatikregeln** ein, z. B. kann im Linkskontext einer Nominalphrase ein Determinator oder ein Zahlwort usw. stehen. Die Kontextdefinition kann auch **textspezifische Charakteristika** erfassen, z. B. kann im Linkskontext einer Nominalphrase eine sich öffnende Klammer stehen, ein Formatierungselement des Typs <*FSU*>, <*KEY*>, <*RES*>, <*HDW*>, <*HDN*> usw.

5.4.4.1.3 Lösungsmöglichkeit 3: Lokale Grammatik

Bei der lexikalischen Analyse identifiziert INTEX die Wortformen, die im Text auftreten. Wortformen sind isoliert gesehen oft ambig, ein Teil von ihnen kann aber durch eine Analyse des Kontexts disambiguiert werden. Der für die Disambiguierung relevante Kontext wird durch eine lokale Grammatik 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: Erkennung von Mehrwortlexemen und Komposita, Repräsentation orthographischer Varianten im Lexikon, Überprüfung der Kongruenz, Identifikation von Zeitangaben usw.

Ein **Transduktor** ist ein endlicher Automat, der zusätzlich eine Ausgabe erzeugt, 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” dient dazu, Substitutionen im Text auszuführen, eine identifizierte Sequenz mit zusätzlicher Information (z. B. einer Wortklasse) zu versehen oder linguistische Markierungen (z. B. die Annotation von Phrasen) in den Text einzufügen.

Beispiel:

Die Darstellung der Disambiguierung der Form *s’* verdeutlicht die Funktionsweise einer lokalen Grammatik[1].

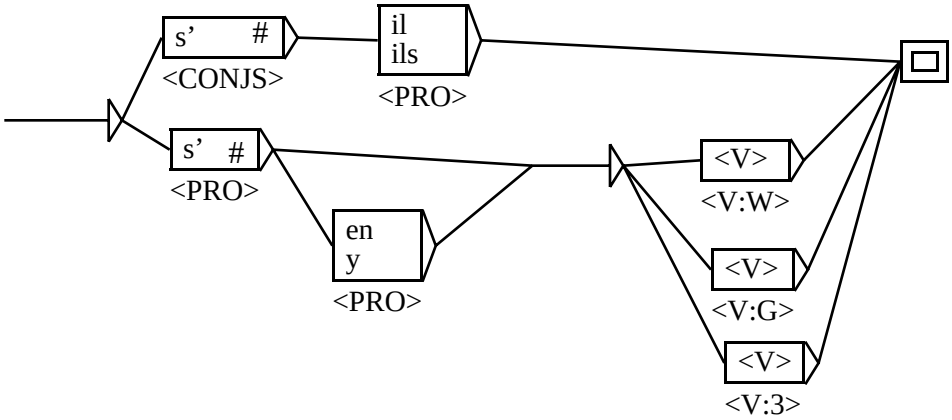
Die Form *s’* kann sowohl ein Pronomen (elidierte Form von *se*) als auch eine Konjunktion (elidierte Form von *si*) sein. Zur Auflösung der Ambiguität werden die unmittelbaren (d. h. lokalen) Kontexte untersucht.

- Die Form *s’* repräsentiert eine Konjunktion nur dann, wenn die Pronomen *il* oder *ils* direkt nachfolgen.
- Die Form *s’* repräsentiert das Reflexivpronomen *se* nur dann, wenn direkt darauf ein Verb folgt, das mit einem Vokal beginnt, oder wenn eines der präverbalen Pronomina *en* oder *y* und ein Verb folgen.

1. Vgl. dazu Silberztein 1993, S. 26.

Die Überprüfung des lokalen Kontexts der Form *s'* genügt, um die Ambiguität aufzulösen.

Abbildung: Transduktor zur Disambiguierung von *s'*



Das Zeichen "#" am Ende eine Knotens bedeutet, daß kein Leerzeichen eingefügt werden darf, der Apostroph muß unmittelbar auf das Wort folgen. <CONJS> steht für die Wortklasse untergeordnete Konjunktion, <PRO> für Pronomen, <V> für Verb. Die Tatsache, daß nur Verben, die mit einem Vokal beginnen, in dieser Struktur auftreten dürfen, ist an dieser Stelle noch nicht repräsentiert.

Lokale Grammatiken bieten sich immer dann an, wenn der Kontext eine Disambiguierung ermöglicht, wie das bei einem Teil der in (siehe Abschnitt 5.4.4.1 Ambiguität, Seite 90) aufgeführten Beispiele der Fall ist. Ein Vorteil von INTEX ist, daß mit den Transduktoren nicht nur ein Formalismus zur Darstellung lokaler Grammatiken, sondern auch automatische Verfahren für ihre Anwendung geboten werden.

5.4.4.2 Vollständigkeit und Struktur der INTEX-Lexika

5.4.4.2.1 Wörter, die nicht im INTEX-Lexikon stehen

Im Durchschnitt werden 5 -10 % der Wortformen eines Dokuments bei der lexikalischen Analyse mit INTEX nicht erkannt und in einer gesonderten Datei ausgegeben. Unter den nicht erkannten Wortformen finden sich:

- Eigennamen: *Blendax, Hoechst*
- Rechtsschreibfehler: *decision*
- Abkürzungen und Akronyme: *BASF, IPC*
- unbekannte Wörter: *périodontique, hydroxy, diastéréoisomères*

Folglich können Nominalphrasen, die eine solche Wortform enthalten, auch nicht erkannt werden, wie beispielsweise *affections périodontiques* (Periodontopathie).

Unbekannte Wörter, die weder Eigennamen noch Rechtschreibfehler oder Abkürzungen oder Akyronyme sind, machen nur ca. 3 % eines Dokuments aus.

5.4.4.2.2 Zusammengesetzte Strukturen

Die differenzierte Analyse der zusammengesetzten Strukturen ist eine der Stärken von INTEX gegenüber vergleichbaren Softwaresystemen. INTEX verfügt über umfangreiche Lexika für nicht-ambige komplexe Strukturen (Sequenzen, die unabhängig vom Kontext immer als Einheit zu behandeln sind) und für ambige komplexe Strukturen (Sequenzen, die in *manchen* Kontexten als eine Einheit zu interpretieren sind).

Die Erkennung komplexer Strukturen sichert, daß eine solche Struktur als Ganzes behandelt wird und verhindert, daß ein Teil oder die Gesamtheit einer solchen Struktur fälschlicherweise als NP-Muster interpretiert wird, wie in

en tout *état de cause* (kein NdeN-Muster),

vorkommend in: [En tout état de cause], il faut bien être conscient du fait que ...

la *forme du chlorure* (kein NdeN-Muster),

vorkommend in: le cation lanthane étant présent [sous la forme du] chlorure ...

au *cours de la procédure* (kein NdeN-Muster),

vorkommend in: [Au cours de la] procédure devant la division d'opposition ...

de *nature cosmétique* (kein NA-Muster)

vorkommend in: Le traitement revendiqué était [de nature cosmétique] et non médicale.

Die Untersuchung des Testkorpus zeigte, daß bis zu 12 % der aus einem Dokument extrahierten Sequenzen komplexe Strukturen enthalten, die allgemeinsprachlich oder

textspezifisch sind. Der Benutzer kann in INTEX sein eigenes Lexikon für komplexe Strukturen anlegen. Komplexe Strukturen, die in einem bestimmten Korpus immer eindeutig zu interpretieren sind, können in einem Präferenzlexikon abgelegt werden.

Im EPA-Korpus gibt es eine Reihe komplexer Strukturen mit adverbialer oder präpositionaler Funktion, die entweder textspezifisch sind oder in der Patentdokumentation häufiger als in allgemeinsprachlichen Texten vorkommen. Ihre Verwendung im EPA-Korpus ist nicht ambig, so daß sie in einem Präferenzlexikon für komplexe Strukturen abgelegt werden. Auf diese Weise werden sie immer als zusammengehörige Einheit behandelt und es wird verhindert, daß ein Teil einer solchen Einheit als Teil eines potentiellen Fachterminus interpretiert wird.

Beispiele:

au désavantage de: z. B. in: au désavantage du demandeur: dem Anmelder zum Nachteil gereichen

de plein droit: von Rechts wegen

dans le cas présent: im vorliegenden Fall

an application de l'article ...: nach Artikel ...

dans un délai de ... mois: innerhalb einer Frist von ... Monaten

5.4.4.2.3 Bindestrichwörter

Bindestriche gelten in INTEX als Separatoren (neben anderen wie Leerzeichen etc.) für einfache Formen. Sie im DELAS aufzuführen wäre ein Widerspruch zur Definition der einfachen Formen. Ein Teil der Bindestrichformen sind im DELAC enthalten. Es sind vor allem Funktionswörter: *celle-ci* (Pronomen), *elle-même* (Pronomen), *ci-après* (Adverb).

Nominale Bindestrichkomposita, z. B. *sous-revendication* (Unteranspruch), *Vice-Président* (Vizepräsident), wurden mit einem Perl-Programm aus dem Text extrahiert und in einem "Benutzerwörterbuch" in INTEX abgespeichert.

5.4.4.3 Komposita der Länge 2 oder Komposita der Länge 3

Dieses Problem wurde zuvor (siehe Abschnitt 5.3.3 Komplexe Nominalphrasen, Seite 77) ausführlich beschrieben. Es kann im Rahmen einer maschinellen Analyse nicht gelöst werden. Im Rahmen einer intellektuellen linguistischen Untersuchung, die hier nicht durchgeführt werden soll, können sicher zuverlässige Interpretationen für einzelne Fälle gewonnen werden.

10-30 % der zweigliedrigen Strukturen des in dieser Studie untersuchten Korpus waren Teil einer mehrgliedrigen Struktur. Dieser Prozentsatz ist zu hoch, um das Problem unberücksichtigt zu lassen. Für die dritte Studie wurde nach der schon erwähnten Heuristik verfahren: Extraktion zweigliedriger und häufig auftretender dreigliedriger Strukturen.

5.5 Studie 3: Anwendung von INTEX

5.5.1 Vorbereitung

Dieser Untersuchung liegt der frz. Teil des Korpus zugrunde, der in der Testreihe 4 zur Evaluierung der Alignierungsalgorithmen verwendet wurde (siehe unter Abschnitt 4.4.2 Testreihen und Ergebnisse, Seite 53). Es handelt sich um sechs kürzere und längere Dokumente aus dem EBK-Korpus, insgesamt ca. 38 000 Wörter. Das kürzeste Dokument ist 140 Sätze lang (5 400 Wörter), das längste Dokument ist 270 Sätze lang (9 800 Wörter).

Für die Bearbeitung mit INTEX wurden einige Vorbereitungen getroffen:

- Erstellung eines Lexikons der unbekannten einfachen Formen
- Erstellung von Präferenzlexika für einfache und komplexe Formen
- Darstellung der zu untersuchenden Strukturen als Automaten
- Bestimmung des optimalen Rechts- und Linkskontexts.

5.5.1.1 Beispiel: Darstellung der Strukturen als Automaten mit Kontextbeschreibungen

Der folgende Automat beschreibt das Bildungsmuster NA.

- Die Kongruenz zwischen Adjektiv, Nomen und Determinator ist explizit durch die Abfolgen der morphosyntaktischen Kategorien auf den verschiedenen Übergangswegen angegeben (z. B. <Det:ms> <N:ms> <A:ms>).
- Statt des Determinators kann ein anderer Linkskontext stehen, der in einem Unterautomaten (*linkskontext*) abgebildet wurde. Dazu gehören verschiedene Arten verbaler Formen (<V:I>, <V:P> etc.), Konjunktionen (<CONJ>, <CONJS>, <CONJC>), Zahlformen (<NB>) etc.. Die Klasse <EPA-Codes> bildet die dokumentspezifischen Formatierungsscodes ab (siehe Abschnitt 3.3.1 Strukturierung eines Dokuments, Seite 28).

Zwischen Determinator und Nomen bzw. zwischen Linkskontext und Nomen können verschiedene Adjektive (z. B. *simple, autre*) stehen, die aufgrund ihrer Semantik für die Bestimmung potentieller Fachtermini unbedeutend sind und deshalb außer Acht gelassen werden können. Sie sind in vier **Unterautomaten** dargestellt: *insertion* (ms), *insertion* (fs), *insertion* (mp) und *insertion* (fp), so daß die Kongruenz zwischen Nomen und eingefügtem Adjektiv berücksichtigt wird (z. B. *insertion* (ms) für die Einfügung von Adjektiven im Maskulinum Singular). Auf das Bildungsmuster folgt der Rechtskontext (z. B. Interpunktionszeichen, Formen konjugierter Verben etc.), dargestellt durch den Unterautomaten *rechtskontext*, auf dessen Abbildung hier verzichtet wurde.

Abbildung: Automat für das Bildungsmuster *NA*

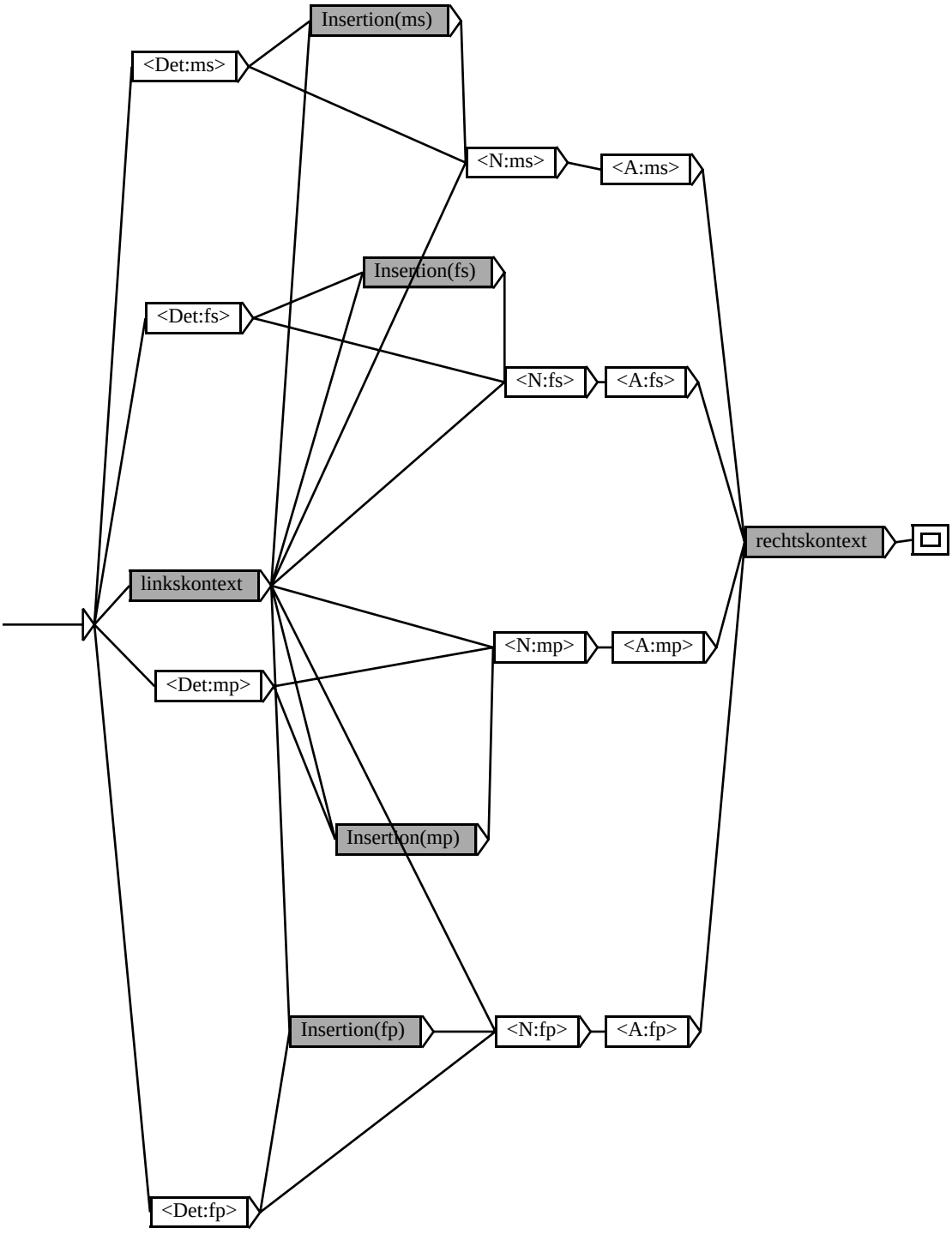


Abbildung: Unterautomat *linkskontext*

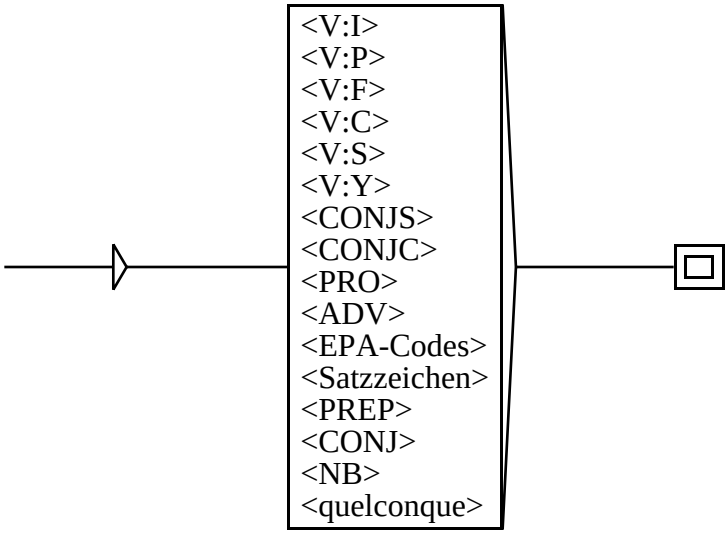
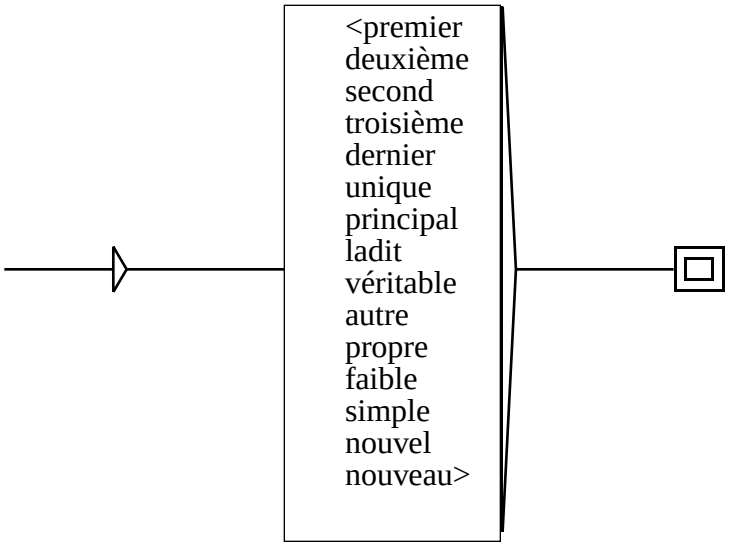


Abbildung: Unterautomat *insertion(ms)*



5.5.2 Ergebnisse

Die Arbeitweise von INTEX und die Vorteile dieses Programmpakets wurden zuvor erläutert. Als Ziel sollte eine möglichst optimale Anwendungweise von INTEX bestimmt werden, in dem Sinne, daß eine größtmögliche Anzahl korrekter Strukturen extrahiert wird (hohe Vollständigkeits- und Genauigkeitsquote).

Alle mit INTEX extrahierten Strukturen wurden intellektuell überprüft. Die Ergebnisse gliedern sich in zwei Gruppen:

- **INTEX-Strukturen**: Strukturen mit Umgebungsdefinition, die bei der Extraktion mit INTEX unter Verwendung der angegebenen Filter eine Korrektheit von mindestens 65 % erreichen. Bei einer Verbesserung der Filter könnte die Qualität der Ergebnisse so weit gesteigert werden, daß eine automatische Extraktion ohne intellektuelle Überprüfung denkbar wird.
- **Ergänzte Strukturen**: Strukturen, deren Extraktion mit INTEX in der jetzigen Form noch keine ausreichende Korrektheit erreicht. Ein zufriedenstellendes Ergebnis kann hier nur mit einer vollständigen intellektuellen Überprüfung erzielt werden.

Die Ergebnisse in den beiden Gruppen werden in einer abschließenden Betrachtung (Gesamtstatistik) zusammengeführt.

5.5.2.1 INTEX-Strukturen

Die Ergebnisse werden für jedes der sechs untersuchten Dokumente gesondert aufgeführt. Für jedes Dokument und für jeden Phrasenbildungstyp werden angegeben:

- die Gesamtzahl der extrahierten Strukturen eines bestimmten Typs (die in der Tabelle mit *gesamt* bezeichnete Spalte).
- die Anzahl der korrekt extrahierten Strukturen (die in der Tabelle mit *korrekt* bezeichnete Spalte).
- die Anzahl der Fälle, in denen eine maximale Nominalphrase von INTEX inkorrekt in Teilphrasen zerlegt wurde (die in der Tabelle mit *falsche Zerlegung* bezeichnete Spalte).

  z. B. NA: *priorité antérieure* vorkommend in *une date de priorité antérieure* (ein früherer Prioritätstag)
- die Anzahl sonstiger Fehler (die in der Tabelle mit *andere Fehler* bezeichnete Spalte), z. B. eine falsche Wortklasse.

  z. B. NdeN: *exempts de sodium*, vorkommend in *les produits chimiques ne sont jamais totalement exempts de sodium* (...daß es natriumfreie Chemikalien nicht gibt.)

Die Zeile *insgesamt* gibt die Summe einer Spalte an.

Tabelle 19. INTEX-Strukturen in Dokument 843

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	130	118	7	5
NdeN	184	172	6	8
NàN	17	14	0	3
NPrepN	19	4	6	9
NN	29	29	0	0
NdeNA	26	22	3	1
NdeNdeN	15	8	3	4
NAA	9	6	0	3
NAdeN	8	8	0	0
NdeNdeNN/A	5	5	0	0
insgesamt	**442**	**386**	**43**	

Tabelle 20. INTEX-Strukturen in Dokument 1089

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	142	112	17	13
NdeN	200	183	20	6
NàN	2	0	1	1
NPrepN	45	31	2	12
NN	1	1	0	0
NdeNA	28	15	13	0
NdeNdeN	32	24	3	5
NAA	6	3	0	3
NAdeN	15	14	1	0
NdeNdeNN/A	5	4	1	0
insgesamt	**476**	**387**	**58**	**40**

Tabelle 21. INTEX-Strukturen in Dokument 2018

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	212	192	0	20
NdeN	130	128	6	6
NàN	5	2	1	2
NPrepN	24	13	4	7
NN	3	3	0	0
NdeNA	57	46	4	7
NdeNdeN	20	17	0	3
NAA	21	16	0	5
NAdeN	3	3	0	0
NdeNdeNN/A	6	4	0	2
insgesamt	**481**	**424**	**15**	**52**

Tabelle 22. INTEX-Strukturen in Dokument 2463

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	161	138	20	3
NdeN	140	125	10	5
NàN	5	0	0	5
NPrepN	20	6	4	10
NN	15	15	0	0
NdeNA	41	36	4	1
NdeNdeN	11	10	1	0
NAA	9	3	6	0
NAdeN	16	14	2	0
NdeNdeNN/A	7	3	3	1
insgesamt	**435**	**350**	**50**	**25**

Tabelle 23. INTEX-Strukturen in Dokument 2653

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	137	90	13	24
NdeN	289	282	7	10
NàN	14	1	0	13
NPrepN	35	13	2	19
NN	0	0	0	0
NdeNA	27	16	9	2
NdeNdeN	36	29	6	1
NAA	7	4	2	1
NAdeN	24	24	0	0
NdeNdeNN/A	3	2	1	0
insgesamt	**572**	**461**	**40**	**70**

Tabelle 24. INTEX-Strukturen in Dokument 2842

Struktur	gesamt	korrekt	falsche Zerlegung	andere Fehler
NA	108	94	6	8
NdeN	121	104	15	2
NàN	4	2	0	2
NPrepN	12	2	3	7
NN	9	9	0	0
NdeNA	23	13	10	0
NdeNdeN	24	16	2	3
NAA	5	2	0	3
NAdeN	4	4	0	0
NdeNdeNN/A	3	1	2	0
insgesamt	**310**	**247**	**38**	**25**

5.5.2.2 Ergänzte Strukturen

Bei der Bearbeitung mit INTEX sollte eine möglichst hohe Korrektheitsquote erzielt werden, was auf Kosten der Vollständigkeit geht. Es hatte sich in der vorgehenden Stu-

die gezeigt, daß bestimmte Strukturen, beispielsweise <N> <N>, auch beim Einsatz der beschriebenen Disambiguierungsstrategien nur mit geringer Korrektheit extrahiert werden können.

Darüber hinaus werden manche potentielle Fachtermini aufgrund der restriktiven Definition der NP-Muster und der Rechts- und Linkskontexte nicht extrahiert. Da eine breitere Kontextdefinition aber zu viele inkorrekte Strukturen mit sich gebracht hätte, wurden diese Fälle gesondert mit INTEX im Text lokalisiert, intellektuell überprüft und in das Gesamtergebnis integriert. Es handelt sich im wesentlichen um folgende Fälle:

1. Adjektiv oder Partizip im Rechtskontext

Viele Formen des Partizip Perfekt Passiv können auch Adjektive sein. INTEX ordnet solchen Formen beide Kategorien zu.

Beispiel:

la littérature spécialisée: Fachliteratur (*spécialisée* ist hier Adjektiv)

un avocat spécialisé dans le droit des brevets: Ein Rechtsanwalt, der sich auf Patentrecht spezialisiert hat (*spécialisé* ist hier Partizip)

Diese Ambiguität kann in der vorliegenden Anwendung nur intellektuell aufgelöst werden. Die folgenden Ausführungen gelten deshalb für Adjektive und für Wortformen, die Adjektiv und Partizip sein können. Präpositionale Nominalphrasen können von Adjektiven und Partizipien eingeleitet werden. Eine korrekte Interpretation erfordert Subkategorisierungsinformationen, wie sie z. B. im System LEXTER (Bourigault 1994) mit Hilfe eines Lernverfahrens integriert werden. Solche Informationen standen hier aber nicht zur Verfügung.

Läßt man Adjektive bzw. Partizipien im Rechtskontext zu, so kommt es zu Zerlegungsproblemen und Überlappungen von NP-Mustern, die vorerst nur intellektuell aufgelöst werden können.

Beispiele:

(1) des impuretés des produits *chimiques* utilisés:Verunreinigungen der verwendeten Chemikalien

richtige Zerlegung:

[des impuretés des [produits chimiques]$_{NA}$]$_{NdeNA}$ utilisés

falsche Zerlegung:

[des impuretés des produits]$_{NdeN}$ chimiques utilisés

(2) la protection *conférée* par le brevet: Schutzbereich des Patents

richtige Zerlegung:

[la protection conférée par le brevet]

falsche Zerlegung:

[la protection conférée]$_{NA}$ par le brevet

2. Determinatoren im Rechtskontext

Die Zerlegung einer maximalen Nominalphrase bestehend aus einem NP-Muster gefolgt von einem Determinator der Gruppe *de, d', de l', du, de la, des* ist wegen der Überlappung der NP-Muster problematisch.

Beispiele:

(1) des impuretés *d'*alcali *des* zéolites: Alkaliverunreinigungen von Zeolithen

richtige Zerlegung:

[des impuretés d'alcali]$_{NdeN}$ des zéolites

falsche Zerlegung:

des impuretés d'[alcali des zéolites]$_{NdeN}$

3. Partizip statt Adjektiv

Es wurde untersucht, ob in den NP-Mustern, die ein Adjektiv enthalten (z. B. NA, NdeNA) statt des Adjektivs auch ein Partizip stehen kann, d. h. eine Wortform, die im INTEX-Lexikon die Wortklasse Partizip, aber nicht die die Wortklasse Adjektiv hat. Es treten dabei ähnliche Zerlegungsprobleme wie in den oben beschriebenen Fällen auf:

Beispiel:

(1) richtig: le brevet *attaqué*: das Streitpatent

(2) richtig: des amines *incorporées:* eingelagerte Amine

aber:

(3) la demande *déposée* ultérieurement

richtige Zerlegung:

[la demande *déposée* ultérieurement]

falsche Zerlegung:

[la demande *déposée*]$_{NA}$ ultérieurement: die Nachanmeldung

(4) une décision *rendue* sur opposition: Einspruchsbeschwerdeverfahren

richtige Zerlegung:

[une décision rendue sur opposition]

falsche Zerlegung:

[une décision rendue]$_{NA}$ sur opposition

5.5.2.3 Sonstige Strukturen

Bei der Durchsicht der Dokumente fielen komplexe Komposita auf, die aufgrund ihrer Frequenz (teilweise bis zu 7 Vorkommen derselben Struktur in einem Dokument) und

ihrer Semantik potentielle Fachtermini sind, deren Bildungsmuster aber nicht automatisch erfaßt werden konnten.

Beispiele:

(1) les connaissances générales de l'homme du métier: allgemeines Fachwissen (der Begriff ist wichtig im Zusammenhang mit der Prüfung der “erfinderischen Tätigkeit”, die ein Kriterium für die Patentfähigkeit einer Anmeldung ist)

(2) la requête en restitutio in integrum: der Antrag auf Wiedereinsetzung /der Wiedereinsetzungsantrag (juristisches Prinzip)

(3) le droit des parties à être entendues: das rechtliche Gehör (juristisches Prinzip)

(4) un sel soluble dans l'eau: ein wasserlösliches Salz (als Synonym zu *un sel hydrosoluble*, das 6 mal im Dok. 2018 vorkommt)

(5) un sel insoluble dans l'eau: wasserunlösliches Salz (kommt 7 mal im Dok. 2018 mal vor)

Vorkommen komplexer Komposita, die durch eine Verkettung derjenigen Strukturen entstanden sind, die in der INTEX-Studie berücksichtigt wurden (wie das Beispiel (1)), sind in der Gesamtstatistik in der mit *Sonstige* bezeichneten Zeile aufgeführt.

5.5.2.4 Gesamtstatistik

Die Ergebnisse der korrekt extrahierten INTEX-Strukturen und der ergänzten Strukturen wurden in einer Gesamtstatistik zusammengefaßt.

Tabelle 25. Gesamtstatistik

Struktur	Summe aus 6 Dok.	Dok. 843	Dok. 1089	Dok. 2018	Dok. 2463	Dok. 2653	Dok. 2842
NA	1065	181	175	252	199	125	133
NdeN	1135	188	214	150	150	305	128
NàN	19	14	0	2	0	1	2
NPrepN	74	4	31	13	6	16	4
NN	69	36	1	3	17	2	10
NdeNA	180	29	23	51	41	20	16
NdeNdeN	112	8	27	19	11	30	17
NAA	48	11	10	16	5	4	2
NAdeN	74	9	14	5	14	27	5
NdeNdeNN/A	19	5	4	4	3	2	1
Sonstige	9	1	7	0	0	0	1
insgesamt	**2804**	**486**	**506**	**515**	**446**	**532**	**319**

5.5.2.5 Fazit

Das INTEX-System erwies sich als effizientes und schnelles Instrument, um die linguistischen Eigenschaften der potentiellen Fachtermini im untersuchten Korpus herauszuarbeiten.

Für das Problem der Ambiguitäten konnten in Teilbereichen Lösungen erarbeitet werden. Die Analyseergebnisse wurden intellektuell überprüft, vor allem auch deswegen, weil der INTEX-Output in der vorliegenden Arbeit den Status eines Zwischenergebnisses hat, das die Grundlage einer kontrastiven (dt.-frz.) Untersuchung ist. Durch die beschriebene Handhabung ist gewährleistet, daß so wenig Fehler wie möglich in die nächste Stufe der Untersuchung eingehen.

Es müßte anhand eines wesentlich größeren Korpus überprüft werden, ob die Frequenz ein brauchbares Kriterium zum Ausfiltern falscher Analysen ist. Ein größeres Korpus könnte auch neue Erkenntnisse für die Lösung des Problems der Vollständigkeits- und der Genauigkeitsquote liefern.

6 Nominale Fachterminologie im Deutschen

Die Bestimmung und Extraktion der Terminologie aus einem dt. Fachtext setzt eine Klärung der formalen Eigenschaften von Fachwörtern voraus. Im folgenden Kapitel werden zuerst einige Arbeiten aus der Forschung zur dt. Fachsprache vorgestellt. Zur Ergänzung werden Ergebnisse eines Forschungsprojekts dargestellt, das sich speziell mit der dt. Patentdokumentation beschäftigte.

Eine dreigliedrige Studie beschreibt anschließend die Extraktion potentieller Fachtermini aus dem EPA-Korpus:

- Studie 1: Linguistische Beschreibung potentieller Fachtermini
- Studie 2: Bestimmung eines geeigneten Werkzeugs
- Studie 3: Ergebnisse.

6.1 Fachsprachenforschung: Formale Eigenschaften der Benennungen im Deutschen

Einige einschlägige Werke zur Terminologie[1]wurden unter dem Aspekt untersucht, wie sie die formalen Eigenschaften der Benennungen darstellen. Dabei fällt auf:

- Diese Fragestellung nimmt in allen Werken nur einen sehr kleinen Raum ein.
- Die Unterschiede von einem Autor zum anderen bei der Darstellung der formalen Eigenschaften der Benennungen sind sehr gering.

Arntz und Picht verweisen in ihrer Definition auf die Ausführungen der Norm DIN 2330 (Begriffe und Benennungen):

> Eine Benennung ist die mindestens ein Wort umfassende Bezeichnung eines Begriffs. Benennungen können Wörter und Wortgruppen sein.
>
> Bei den Wörtern sind zu unterscheiden: Stammwörter (z. B. “Licht”), zusammengesetzte Wörter (z. B. “Glüh/lampe/n/fass/ung/en”), und abgeleitete Wörter (z. B. “Ver/bind/ung”); eine Wortgruppe besteht aus mindestens zwei getrennt geschriebenen, syntaktisch verbundenen Wörtern (z. B. “zulässige Abweichung”)[2].

Die Norm erwähnt auch die formalen Anforderungen an Benennungen:

> Benennungen sollen sich zwanglos in das Sprachgefüge einordnen.
>
> Die Benennungen sollen sein:
>
> - angemessen kurz

1. Fluck 1985; Hoffmann 1985; Drozd & Seibicke 1973; Arntz & Picht 1982; Reinhart et al. 1992.
2. Arntz & Picht 1982, S. 109.

- einprägsam
- leicht sprechbar
- geeignet zum Bilden von Ableitungen[1].

“Die wichtigste Wortart unter den spezialsprachlichen lexikalischen Einheiten bilden die Substantive[2]”. In der Beschreibung der “äußerlichen“ Eigenschaften der Benennungen beziehen sich die Autoren deshalb in erster Linie auf Nomen. Sie unterscheiden zwischen der fachbezogenenVerwendung gemeinsprachlicher Wörter (*Terminologisierung*) und der Neubildung von Benennungen.

6.1.1 Terminologisierung

Grundsätzlich kann jede Bezeichnung des Allgemeinwortschatzes terminologisiert werden, d. h. ihr wird in einer Fachsprache (oder in mehreren) eine ganz bestimmte Funktion zugewiesen. Ein Beispiel ist das gemeinsprachliche Wort “Wurzel”, dem in der Mathematik und in der Zahnmedizin neue spezifische Bedeutungen zugewiesen worden sind. Eine solche Übertragung von Bedeutungen beruht meist auf der Erkenntnis von Ähnlichkeiten. Das wird besonders deutlich dort, wo die Benennungen menschlicher Körperteile auf Teile von Maschinen,Werkzeugen usw. übertragen werden (z. B. Kopf, Nase, Zahn, frz.: *tête*, *nez*, *dent*). Zu dieser Gruppe gehören auch solche Fachwörter, die ihre Grundlage in der definitorischen Festlegung gemeinsprachlicher Wörter haben. Dies gilt beispielsweise für viele Grundbegriffe der Physik: Raum, Zeit, Bewegung (frz.: *espace*, *temps*, *mouvement*).

Unter Einhaltung der Einheit von Form und Inhalt des Fachworts müssen terminologisierte Wörter als neue Termini betrachtet werden, selbst wenn dabei keine neue Lautform entsteht. Drozd und Seibicke verweisen auf spezielle Eigenschaften, die das terminologisierte Wort im Gegensatz zu dem allgemeinsprachlichen Wort hat, das als Ausgangspunkt für die Terminologisierung diente: z. B. die Fähigkeit zur funktionellen Pluralbildung (wie bei: *Sande, Öle, Fette*), funktionelle Veränderungen in der Flexion (wie bei: *Dorne* anstatt *Dörner* bzw. *Dornen*) oder im Genus (das *Ersatzteil*, das *Vorsatz*, das *Filter*)[3].

6.1.2 Wortzusammensetzung

Durch die Wortzusammensetzung (Kompositabildung) entsteht einer der produktivsten Typen des Terminus. Das Substantiv stellt den wichtigsten Bestandteil von Wortzusammensetzungen dar. Die Fähigkeit des Substantivs, im Grundwort den Grundbegriff und im Bestimmungswort die Merkmalseinschränkung darzustellen, wird in allen Fachsprachen reichlich genutzt. Die Anzahl von Bestandteilen einer

1. DIN 2330, zitiert nach Arntz & Picht 1982, S. 109.
2. Fluck 1985, S. 38.
3. Beispiele aus: Drozd & Seibicke 1973, S. 147.

Wortzusammensetzung - nicht nur von Substantiven, sondern auch von anderen Wortarten - ist relativ unbeschränkt, aber oft durch sprachpsychologische Aspekte bestimmt. Drozd und Seibicke zitieren die Häufigkeitsuntersuchungen von Ischreyt, nach denen in einer Grundgesamtheit von 11 500 Termini zweigliedrige Komposita einen Anteil von 52 % ausmachen, dreigliedrige einen Anteil von 28 % und die vier- und mehrgliedrigen nur noch einen Anteil von 6 % haben[1].

Die Untersuchungen fachsprachlicher Komposita beziehen sich auf die Wortarten der einzelnen Konstituenten, auf deren Anordung (Fugenformen etc.) und die semantischen Relationen, die zwischen ihnen vorliegen. Letztere werden immer im Zusammenhang mit den Motivationsmodellen für Fachtermini und allgemeinen Grundsätzen der Terminologie betrachtet[2]. Während die Untersuchung der Wortarten der Konstituenten, von geringfügigen Unterschieden von einem Autor zum anderen abgesehen, relativ gleichbleibend ist, wird die Analyse der Beziehungen zwischen den Konstituenten recht unterschiedlich dargestellt. Es wurde die Arbeit von Reinhardt et al.[3] herausgegriffen, da sie mit relativ wenig Hintergrundwissen zu verstehen ist. Ihre Klassifikation der Zusammensetzung ist aber in einigen Punkten genauso fraglich wie die anderen Darstellungen zu diesem Thema.

Reinhardt et al. unterscheiden entsprechend ihrer Einteilung der Motivationsmodelle vier Arten von Zusammensetzungen.

1. Die motivierend-determinierende Zusammensetzung

Hier dient die dem Ausgangswort vorangestellte Konstituente zur Differenzierung, die meistens einer begrifflichen Einengung gleichkommt. Da die grammatischen Beziehungen zwischen den Konstituenten eines Kompositums normalerweise nicht mehr ersichtlich sind, können die inhaltlichen Beziehungen nur noch aus dem Kontext oder infolge Sachkenntnis abgelesen werden.

z. B. Schaumglas, Reineisen, Außenantenne

2. Die präzisierende Zusammensetzung

Die dem Ausgangswort nachgestellte Konstituente dient der Präzisierung des im Ausgangswort Genannten hinsichtlich der Größe, der Beschaffenheit, der Form usw. Bei Vorgangsbenennungen kann auf diese Weise präzisiert werden, ob der Vorgang in einer bestimmten Phase, in seiner Gesamtheit oder in einer bestimmten Abstraktionsstufe gesehen werden soll.

z. B. Achsensystem (als System angeordnete Achsen), Schleifbeginn, Schleifprozeß

1. Drozd & Seibicke 1973, S. 147.

2. Vgl. dazu Felber und Budin 1989, Kap. 3; Fluck 1985, Kap. 6; Drozd & Seibicke 1973, Kap. II.2 und Kap. IV.3; Arntz & Picht 1982, Kap.3.

3. Reinhardt et al. 1992.

3. Die kategorisierende Zusammensetzung

Hier dient die nachgestellte Konstituente in erster Linie dazu, die Benennung einer der in der Technik relevanten Kategorien - Mittel, Vorgang, Objekt - zuzuordnen. Muß beispielsweise in einem Fachtext, der Probleme des Messens behandelt, ein technisches Mittel zur Ausführung von Meßvorgängen allgemein benannt werden, so kann ohne weiteres die Benennung *Meßmittel* gebildet werden; es sind aber auch andere Benennungen wie *Meßgerät*, *Meßvorrichtung*, *Meßeinrichtung*, *Meßsystem* usw. möglich. Stets wird ein Mittel zum Messen benannt, wobei aber durch die Bedeutung des gewählten Elements eine nähere Spezifizierung, eine Präzisierung erfolgt. Die allgemeinsten dieser Elemente bringen neben ihrer Kategorisierungswirkung nur wenig präzisierende Bedeutung ein und nähern sich damit den Suffixen.

4. Die komprimierende Zusammensetzung

Hier geht es nicht um eine Terminusbildung wie in den drei bisher dargestellten Verfahren, sondern um eine äußerste Komprimierung syntaktischer Konstruktionen. Sie können allerdings bei häufigem Auftreten und der Herausbildung entsprechender Verhältnisse in der objektiven Realität zu Benennungen werden. Bei der Bestimmung vorhandener Bildungen ist deshalb eine eindeutige Zuordnung oft schwierig. In den meisten Fällen handelt es sich um eine Konstruktion aus einer substantivischen Vorgangsbenennung und dem grammatischen Objekt; als vollwertiges Äquivalent ist in der Regel eine Konstruktion mit einem genitivischen Attribut einsetzbar.

z. B. Temperaturerhöhung (Erhöhung der Temperatur), Werkzeugherstellung (Herstellung von Werkzeugen).

6.1.3 Wortableitungen

Selbst wenn die Wortzusammensetzung die typischste Bildungsweise von Termini in dt. Fachsprachen darstellt, so ist die Wortableitung im Deutschen dennoch als eine Wortbildungsart zu behandeln, die in sämtlichen Funktionssprachen und bei den grundlegenden Wortarten hohe Produktivität aufweist.

Im Gegensatz zur Zusammensetzung handelt es sich hier um Bildungen mit Hilfe von Wortbildungsmorphemen (mit Präfixen oder Suffixen) bzw. um Ableitungen ohne “sichtbare” Morpheme (z. B. *Belag* von *belegen*, *Griff* von *greifen*). In einigen Fällen bereitet die Abgrenzung der Präfixe und der Suffixe von gleichlautenden freien Morphemen Schwierigkeiten, z. T. handelt es sich um fließende Übergänge.

Drozd und Seibicke[1] weisen speziell auf *die* Leistungen der Wortableitung und der Präfigierung hin, die sie als terminologisch markierte Bildungsweisen einstufen. In ihrer Aufzählung zeigen sie, in welchen Bereichen der Wortableitung sich fachsprachliche Leistungen besonders produktiv bemerkbar machen. Sie zählen dazu:

1. Drozd & Seibicke 1973, Kap. IV, 4.1.3.

1. das Suffix *-er* in Personen- und Gerätebezeichnungen:

   z. B. Dreher, Weber, Bohrer, Rechner

2. das Suffix *-ling* zur Bezeichnung von Tieren und Pflanzen und in der Metallverarbeitung

   z. B. Engerling, Setzling, Rohling

3. das Suffix *-ung* für ablaufende und abgeschlossene Vorgänge

   z. B. Kühlung, Zeichnung, Formung

4. das Suffix *-heit* zur Bildung von substantivischen Eigenschaftsbezeichnungen

   z. B. Echtheit, Feinheit, Korrektheit

5. das Suffix *-keit* zur Bildung von substantivischen Eigenschaftsbezeichnungen (es hat dieselbe Bedeutung wie das Suffix -heit, wird aber auf andere Basen angewandt)

   z. B. Feuerbeständigkeit, Zähigkeit, Schnelligkeit

6. das Suffix *-bar*, das vor allem mit einem passivfähigen Verb als Basis verknüpft wird

   z. B. waschbar, entzündbar, konstruierbar

7. Suffixe zum Ausdruck der Negation und der Gegensätzlichkeit

   z. B. Mißernte, unedel (Metall), Nichtleiter

6.1.4 Konversion, Entlehnung, Kürzungsverfahren

Unter **Konversion** versteht man den Übergang von Wörtern aus einer Wortklassse in eine andere[1]. Am bekanntesten und zugleich am produktivsten ist hier die Substantivierung, die neue terminologische Einheiten schafft: das *Schmelzen*, das *Flachschrägwalzen*. Neben den Infinitiven ist die Konversion noch bei Namen produktiv. Konvertierte Namen sind häufig in den naturwissenschaftlichen Bereichen Chemie und Physik, treten aber auch in anderen Fächern auf: *Celsius*, *Hertz*, *Röntgen*, *Zeppelin*. Solche Namen können weiter zu Adjektiven und Verben konvertiert (*galvanisch*, *röntgen*, *pasteurisieren*) oder mit Suffixen versehen werden (z. B. das chemische Element *Einsteinium*).

Von **Entlehnung** kann man dann sprechen, wenn ein Terminus aus einer fremden Sprache unter Anpassung an das morphologisch-phonologische System in die Empfängersprache übernommen wurde: *Input*, *Diagnose*, *Software*. Entlehnungen größeren Ausmaßes werden u. a. dann vorgenommen, wenn technische Neuerungen oder

1. Reinhardt et al. 1992 ordnen die Konversion unter dem Abschnitt “Präfixbildung, Suffixbildung, implizite Ableitung” (S. 22) ein.

wissenschaftliche Erkenntnisse aus einem fremden Land übernommen werden (z. B. in der Datenverarbeitung). Im Gegensatz zur Entlehnung überträgt die Lehnübersetzung die einzelnen Wortelemente in die eigene Sprache, ohne die Struktur der Benennung zu verändern, z. B. *Luftbild* (aus *air photo*), *Flutlicht* (aus *flood light*), *chemin de fer* (aus *Eisenbahn*).

Wesentliches Kennzeichen fachsprachlicher Kommunikation ist auch die **Abkürzung**. Mit ihrer Hilfe werden mehrgliedrige Wörter am Anfang, in der Mitte oder am Ende gekürzt, (*Lok* aus *Lokomotive, Bus* aus *Autobus* bzw. *Omnibus*, *Krad* aus *Kraftrad*), zusammengezogen (*Radar* aus *radio detection and ranging*) oder durch Buchstabenwörter ersetzt (*EDV* aus *Elektronische Datenverarbeitung*, *Pkw* aus *Personenkraftwagen*).

6.1.5 Wortgruppen

Ein Terminus, der aus mindestens zwei getrennt geschriebenen, syntaktisch verbundenen Wörtern besteht, wird von Reinhardt et al. als *Wortgruppe*, von Arntz und Picht als *Wortzusammenstellung*, von Fluck als *Mehrwortbenennung* und von Bergenholtz und Tarp[1] als *Mehrwortterminus* bezeichnet.

Es ist auffallend, daß Wortgruppen von einigen Autoren gar nicht oder nur sehr kurz erwähnt werden, und z. T. nur als eine spezielle Realisierung eines Kompositums betrachtet werden, quasi als Übergangszustand bis zur Bildung eines geeigneten Terminus. Drozd und Seibicke schreiben dazu:

> Die Varianz zwischen externen Syntagmen und terminologisierten Einheiten macht sich im Differenzierungsprozeß zwischen den Nichtfachsprachen und den Fach- und Wissenschaftssprachen allgemein geltend:
>
> externe Syntagmen: trockene Masse, buntes Metall, hoher Ofen
>
> fachsprachliche Benennungsstrukturen: Trockenmasse, Buntmetall, Hochofen
>
> ... Die Terminologielehre interessiert sich für die Veränderung der onomasiologischen Struktur, die sich beim Wandel von der Wortgruppe zur Wortzusammensetzung vollzieht[2].

Fluck nennt Gründe, warum ein Kompositum einer Wortgruppe vorzuziehen ist:

> Die Zusammensetzung wird in der Fachsprache der Naturwissenschaft und der Technik deshalb gerne verwendet, da man mit ihrer Hilfe das Ziel der Sprachökonomie erreichen kann. Einerseits wird der sonst durch längere und umständlichere Konstruktionen (z. B. präpositionale Fügungen, Attribute, Appositionen, Relativsätze usw.) zu umschreibende Begriff durch die Zusammensetzung in einer

1. Bergenholtz & Tarp 1994.
2. Drozd & Seibicke 1973, S. 131/132.

> Benennung ökonomisch zusammengefaßt. Andererseits wird die Flexion für die nominalen Komposita vereinfacht. Denn es wird nur noch der zweite Bestandteil des Kompositums flektiert, so daß diese Art der Mehrwortbenennung auch dann noch sprachökonomischer ist, wenn sie sich von der Wortgruppe im äußeren Umfang kaum unterscheidet.[1]

Reinhardt et al. betrachten Wortgruppen (dort auch *Wortgruppenlexeme* genannt) als Spezialfall von Komposita. Der Vorteil der Wortgruppenlexeme gegenüber der Zusammensetzung besteht ihrer Ansicht nach darin, daß die Beziehungen zwischen den Konstituenten noch durch besondere Elemente ausgedrückt werden, was eine sehr ausführliche Motivation ermöglicht. Sie erwähnen folgende Bildungsverfahren:

- Substantiv + Adjektiv bzw. Partizip

  z. B.: kinetische Energie, feuerhemmender Baustoff

- Substantiv + nachgestellte präpositionale Fügung

  z. B.: Schleifen von Hand, Füllstück für Kommutatorfahne

6.1.6 Fazit

In allen untersuchten Arbeiten zur dt. Terminologie wird der Fachterminus in erster Linie oder sogar ausschließlich als Nomen bzw. Nominalkompositum charakterisiert. Mehrworttermini spielen eine untergeordnete Rolle. Die Ausgangslage gestaltet sich somit ganz anders als im frz. Korpus, in dem, bedingt durch die Wortbildungseigenschaften des Französischen, Mehrwortverbindungen im Vordergrund stehen.

Die Extraktion der potentiellen Fachtermini aus dem dt. Teil des EPA-Korpus muß den Schwerpunkt auf Nomen bzw. Nominalkomposita legen. Es ist zu prüfen, welche Rolle Mehrworttermini in diesem Korpus spielen.

6.2 Das PADOK-Projekt

6.2.1 Zielsetzung

Das Projekt PADOK (Patentdokumentation) hatte zum Ziel, für ein Deutsches Patent- und Fachinformationssystem (DPI) die geeignetste maschinelle Inhaltserschließungskomponente durch Tests und Bewertungen herauszufinden. Es wurden vier Systeme zur Inhaltserschließung getestet. Kernpunkt der PADOK-Bewertung war ein möglichst realitätsnaher Retrievaltest auf der Basis von mindestens 10 000 Dokumenten. Die Inhaltserschließung dieser Dokumentmenge sollte die Grundlage für den Aufbau von vier Testdatenbanken sein.

1. Fluck 1984.

Die Inhaltserschließung extrahiert Deskriptoren aus den Dokumenten. Deskriptoren sind entweder Einzelwörter oder Mehrwortverbindungen aus einem Dokument, von denen angenommen wird, daß sie nach Anwendung verschiedener Gewichtungsverfahren den Inhalt eines Dokuments insoweit beschreiben, daß sie als Suchwörter in einer Recherche verwendet werden können. Die von den verschiedenen Systemen erschlossenen Deskriptoren wurden in eine Testdatenbank eingebracht, die durch Retrievaltests evaluiert wurde. Die Ergebnisse lassen Rückschlüsse auf die Qualität der Inhaltserschließung zu.

Es darf angenommen werden, daß die aus einem Dokument gewonnenen Deskriptoren zumindest einen Teil der Fachterminologie dieses Dokuments abdecken. Unter diesem Aspekt ist aus dem PADOK-Projekt insbesondere die folgende Fragestellung interessant:

- Welche Arten von Deskriptoren werden von den Inhaltserschließungskomponenten aus den Dokumenten gewonnen?
- Wie ist die Qualität der verschiedenen Arten von Deskriptoren?

Im PADOK-Projekt standen vier Inhaltserschließungsvarianten (bzw. Systeme) zum Test zur Verfügung:

1. Freitext

   Der Text wird in einzelne Wortformen (Zeichenfolge zwischen Separatoren) zerlegt. Kontextoperatoren (z. B. zwei Deskriptoren sollen im gleichen Satz vorkommen) sind beim Retrieval möglich.

2. PASSAT

   Wortformen werden auf ihre Grundformen zurückgeführt, Komposita in ihre Grundbestandteile zerlegt.

3. CTX

   Im Gegensatz zu PASSAT können Teilelemente von Komposita von Einfachdeskriptoren getrennt werden. Zusätzlich werden aus den Nominalphrasen zweigliedrige Mehrwortbegriffe (komplexe Deskriptoren) extrahiert. Kontextoperatoren sind eingeschränkt möglich.

4. DETECT

   DETECT extrahiert Nominalphrasen maximaler Länge (nicht wie bei CTX zweigliedrige Relationen)

Es steht außer Frage, daß Fachterminologie vor allem in der Form von Nomen und Nominalkomposita auftritt, d. h. - in der Terminologie des Information Retrievals - als Einfachdeskriptoren. Für die vorliegende Arbeit ist die Frage interessant, welche Art von Mehrwortdeskriptoren in der Patentdokumentation berücksichtigt werden sollten. Deshalb sind hier aus der PADOK-Studie in erster Linie die Systeme CTX und

DETECT relevant, da sie komplexe Deskriptoren aus dem Text gewinnen, d. h. Begriffsbenennungen, für die in einer Sprache keine einwortigen lexikalisierten Bezeichnungen existieren. Bei CTX ist im Gegensatz zu DETECT die Art der Mehrwortbegriffe (Maximalausdehnung vs. Zweiwortpaare) thematisiert.

6.2.2 Das System CTX

Als System zur Erschließung von Texten für Informationssysteme verarbeitet CTX Texte in natürlicher Sprache[1]. Bei der Verarbeitung werden von CTX die folgenden Funktionen ausgeführt:

1. Die laufenden Wörter der Texte werden unterschieden nach dem Kriterium ihrer Relevanz als Deskriptoren, nur Wortformen der Wortklassen Substantiv, Verb und Adjektiv werden zur Deskribierung der Texte herangezogen. Wortformen anderer Wortklassen werden als Stopwörter ausgefiltert.

2. Die deskriptorrelevanten Wortformen werden auf ihre Grundform zurückgeführt und stellen die Einfachdeskriptoren dar.

3. Den im Text auftretenden Komposita werden über Thesaurus sinnvolle Teilwörter als zusätzliche Deskriptoren zugeordnet; über denselben Thesaurus werden auch zu entsprechenden Deskriptoren morphologisch und semantisch verwandte Grundformen zugeordnet.

4. Die Nominalgruppen im Text werden syntaktisch analysiert und in Form zweistelliger Relationen als Deskriptoren zur Verfügung gestellt. Die folgenden Nominalgruppenrelationen werden dabei berücksichtigt:

- **A-Relation**: attributives Adjektiv + Substantiv
- **G-Relation**: Substantiv + substantivisches Genitivattribut
- **P-Relation**: Substantiv + substantivisches Präpositionalattribut
- **K-Relation**: zwei koordinierte Substantive

Zur Erzeugung der Deskriptoren führt CTX (im Gegensatz zu DETECT) eine vollständige Syntaxanalyse auf der Basis mehrerer Wörterbücher aus (morphosyntaktisches Wörterbuch SADAW, Derivations- und Zerlegungswörterbuch etc.).

Beispiel einer Texterschließung durch CTX[2]

Text:

> Zur Begrenzung des Wärmeflusses in einem doppelwandigen Rohr wird dieses derart ausgebildet, daß eine bestimmte vorgegebene

1. Vgl. dazu: Womser-Hacker 1986, S. 38 f. und Zimmermann et al. 1983.
2. Aus: Krause 1987, S. 57 f.

> Temperaturdifferenz zwischen der äußeren Wandung und der inneren Wandung eine Trennung der Wandungen voneinander bewirkt, wobei in den entsprechenden Hohlraum ein Gasgemisch eingeleitet wird, um den Wärmeübergangswiderstand zu erhöhen.

Einfachdeskriptoren (in der Reihenfolge ihres Auftretens im Text):

Begrenzung, Wärmefluß, doppelwandig, Rohr, ausbilden, bestimmt, vorgegeben, Temperaturdifferenz, äußer, Wandung, inner, Wandung, bewirken, einstehen, Hohlraum, Gasgemisch, einleiten, Wärmeübergangswiderstand, erhöhen.

Komplexe Deskriptoren:

1. A-Relation:

   doppelwandiges Rohr, vorgegebene Temperaturdifferenz, äußere Wandung, innere Wandung

2. G-Relation:

   Begrenzung Wärmefluß, Trennung Wandung

3. P-Relation:

   Wärmefluss Rohr (aus: zur Begrenzung des Wärmeflusses in einem doppelwandigen Rohr)

   Temperaturdifferenz Wandung (aus: eine bestimmte vorgegebene Temperaturdifferenz zwischen der äußeren Wandung und der inneren Wandung).

6.2.3 Das System DETECT

In den herkömmlichen Freitext-Retrievalsystemen stehen keine komplexen Deskriptoren zur Verfügung. Bei der Recherche kann dieses Manko durch die Verwendung mehrerer Deskriptoren zusammen mit Kontextoperatoren ausgeglichen werden[1]. Ausgangspunkt für DETECT[2] war die Absicht, einen Kontextoperator “Nominalphrase” als Ersatz oder Ergänzung der üblichen Wortabstandsoperatoren zur Verfügung zu stellen und zu evaluieren.

DETECT ist eine Programmierumgebung für die Entwicklung und den Test kontextfreier Analysegrammatiken sowie deren Abarbeitung. Das Parseprogramm arbeitet mit dem Ansatz des partiellen Parsings (“Longest-Match-Verfahren”) anstelle des vollständigen Parsings[3]. Partielles Parsing als eine spezielle Technik der maschinellen Syntaxanalyse heißt hier, daß nicht jedes Wort eines Eingabesatzes der gleichen Analyseprozedur unterzogen wird. Eine vollständige Syntaxanalyse ist für den angestreb-

1. Z. B. *WITH*: Vorkommen im gleichen Satz, *NEAR x*: Wort_1 ist maximal *x* Wörter entfernt von Wort_2.
2. Beschreibung des Systems in: Krause 1987.
3. Beschreibung der Methode u. a. in Seelbach 1975.

ten Zweck nicht notwendig. Nominalphrasen stellen linguistisch gesehen autonome Bestandteile eines Satzes dar. Sie treten als kontinuierliche Wortgruppen auf und besitzen eine relativ stark kanonisierte Binnenstruktur. Problematisch ist lediglich der Bereich der sie umgebenden Satzstruktur, der bei der partiellen Analyse in einigen Fällen zu systematischen Fehlinterpretationen führen kann. Dies ist z. B. immer dann der Fall, wenn der Unterschied zwischen freier Angabe und Objektergänzung oberflächensyntaktisch nicht markiert ist, sondern nur durch den Einbezug der Verb- und Präpositionalvalenzen sowie entsprechender Tiefenkasusanalyse der nominalen Aktanten semantisch ermittelt werden kann.

Die von DETECT verwendete Backus-Naur ähnliche Notation mit der Möglichkeit der Formulierung nicht attribuierter Ersetzungsregeln legt die Konzeption einer Phrasenstrukturgrammatik für die Analyseregeln nahe. Das läßt sich in der Praxis aber nicht immer streng durchführen, da dafür zum einen nicht alle erforderlichen Sprachkonzepte in DETECT vorhanden sind (z. B. Linksrekursion) und zum anderen ein linguistisch “reiner” Ansatz in jedem Fall auf ein spezielles Analyselexikon und die Möglichkeit des Einbringens von prozeduralen Restriktionsregeln angewiesen wäre. DETECT stellt demgegenüber einen stark erweiterten Ansatz des Patternmatchings dar, dessen Besonderheiten eine eigene Methodik für die Grammatikentwicklung und die syntaktische Analyse verlangen.

6.2.4 Vergleichende Evaluierung: DETECT vs CTX

Die Evaluierung in PADOK kann Hinweise geben auf die Leistungsfähigkeit eines Analysesystems auf der Grundlage des partiellen Parsings (DETECT) im Vergleich zu Lösungen mit vollständiger Syntaxanalyse (CTX).

Probleme bei der Bewertung bereitet nicht nur der unterschiedliche Ansatz und der unterschiedliche Leistungsumfang der beiden Systeme, sondern auch der unterschiedliche Entwicklungsstand.

Leistungsumfang

Das Ziel von DETECT ist die Analyse von Nominalphrasen maximaler Länge. Die interne Struktur dieser Nominalgruppen wird nicht ausgegeben. In ihrem Umfang entsprechen sie in etwa den syntaktischen Relationen von CTX.

Beiden Systemen ist gemeinsam, daß alle Elemente des Satzes daraufhin untersucht werden müssen, ob sie Elemente der Zielstruktur Nominalgruppe darstellen. Beide Systeme unterliegen der Restriktion, daß sie koordinierte Nominalgruppen nur erkennen können, wenn sie als kontinuierliche Sequenz auftreten, d. h. nicht von Attributen unterbrochen werden.

Als wesentlicher Unterschied bleibt jedoch der Rekurs auf die notwendigen Resourcen. Die Grammatik von DETECT basiert auf begrenzten Listen von Endungen und Funktionswörtern, deren Inventar nach der Anpassungsphase als abgeschlossen

gelten kann. Demgegenüber arbeitet CTX mit einem umfangreichen Lexikon, das den gesamten Wortsschatz der zu verarbeitenden Texte abdecken muß und dazu morphologische und syntaktische Informationen enthält. Bei der Verarbeitung wird mit Hilfe dieses Lexikons in erster Linie die Reduktion aller Textwortformen durchgeführt; dies zählt jedoch nicht zum Leistungsumfang von DETECT.

CTX

Die Syntaxanalyse von CTX wird zur Aufstellung von Mehrwortdeskriptoren genutzt. Dabei werden komplexe Nominalphrasen auf Vorkommen bestimmter Relationen (A-Relation, G-Relation, P-Relation, K-Relation) untersucht. CTX baut diese Relationen grundsätzlich zweistellig auf, d. h. im Text vorliegende umfangreichere syntaktische Relationen werden in zweistellige Relationen der entsprechenden Einfachdeskriptoren aufgelöst. Den komplexen Deskriptoren ist mit Ausnahme der Adjektiv-Relation nicht mehr zu entnehmen, welche Relation ihnen zugrundelag.

Grundlage für die Bewertung ist der Text der Dokumente. Es wurde zuerst intellektuell ermittelt, welche Mehrwortdeskriptoren Ergebnisse der Analyse von CTX sein müßten. Untersucht wurde eine Textmenge von ca. 140 Dokumenten (Überschrift und Abstract). Von 2700 in den Dokumenten existierenden Relationen wurden 72 % richtig analysiert[1]. Die Ergebnisse zeigen, daß die Syntaxanalyse im Bereich der A-Relation (82 % richtig) und der G-Relation (88 % richtig) durchaus zufriedenstellende Korrektheit erreicht. Problematisch erscheint die P-Relation, die lediglich knapp die Hälfte (43 %) der angestrebten Mehrwortdeskriptoren richtig analysieren kann. Die K-Relation, die etwa zwei Drittel (63 %) der Mehrwortdeskriptoren richtig ermittelt, wirkt ebenfalls nicht ganz überzeugend.

DETECT

Aus 250 von DETECT analysierten Dokumenten sind von 5200 maximalen Nominalgruppen 66 % richtig analysiert. Die maximalen Nominalgruppen aus DETECT enthalten potentiell eine ganze Reihe zweistelliger Relationen. So lassen sich die von CTX und DETECT ermittelten Ergebnisse nicht direkt vergleichen, es läßt sich nur eine Trendaussage ableiten: Da die maximalen Nominalphrasen i. a. mehrere CTX-Relationen enthalten und auch die nicht völlig korrekt analysierten Nominalgruppen richtig analysierte Teile enthalten, können die Zahlen als Hinweis darauf interpretiert werden, daß mit DETECT eine ähnlich hohe Erfolgsquote wie mit CTX erreichbar ist.

Die Analyse syntaktischer Relationen im Bereich der Nominalgruppe wird von CTX zu ca. 75 % erreicht. Demgegenüber sind über 90 % der zweistelligen syntaktischen Relationen korrekt innerhalb einer von DETECT analysierten Nominalgruppe zu fnden. Dieses Verhältnis ist zu relativieren durch ca. 500 Fälle, in denen die maximale Nominalgruppe nicht analysiert wurde. Die syntaktische Analyseleistung von CTX

1. Die genauen Ergebnisse sind in Krause 1987, S. 124 ff. nachzulesen.

und DETECT liegt somit im gleichen Bereich. Partielles Parsing ist eine ernstzunehmende Alternative zu einer vollständigen Syntaxanalyse.

Die Ergebnisse von Testläufen führten im PADOK-Projekt zu einer Abweichung von der ursprünglichen Planung. Vortests hatten ergeben, daß die Patenttexte extrem lange Nominalphrasen enthalten, die nicht problemlos in eine Datenbank von Deskriptoren integriert werden können. Vollständige Nominalphrasen als Suchhilfe werden erst dadurch sinnvoll, daß der Benutzer sie während der Recherche durchliest (z. B. alle Nominalphrasen, in denen der Deskriptor *x* vorkommt) und die seiner Suchintention entsprechenden ankreuzt. Die Nominalphrasen klären somit primär die Rechercheabsicht. Es ist aber, abgesehen von Einzelfällen, nicht damit zu rechnen, daß ohne vorheriges Anbieten solcher Listen komplexe Nominalphrasen als Suchbegriffe effizient nutzbar gemacht werden können. In diesem Fall dürfen die Nominalphrasenlisten, die der Benutzer durchzusehen hat, nicht zu lang und unübersichtlich sein. Dieser Fall lag aber bei den PADOK-Texten vor mit einer Textstruktur, die fast nur überlange komplexe Nominalphrasen kennt. Als einzig verbleibender Nutzen von DETECT wäre somit eine Erweiterung der Kontextoperationen geblieben. Der Benutzer hätte statt im “gleichen Satz” in der “gleichen Nominalphrase” suchen können. Dieser Vorteil schien aber zu gering, um die aufwendigen Retrievaltest zu rechtfertigen. Die Gründe für den Rückzug von DETECT lassen sich jedoch nicht generalisieren. “Für die Patentdokumente müßte DETECT die zugrundeliegende Grammatik so weit verändern, daß kürzere Phrasen entstehen (Einschränkung des Maximalansatzes)”.[1]

Vorteile komplexer Deskriptoren für die Recherche

Im Rahmen der Retrievaltests wurden die Anfragen mit komplexen Deskriptoren in Suchanfragen mit zwei Einfachdeskriptoren mit und ohne Satzkontext nachgestellt. Bei Verwendung zweier Einfachdeskriptoren ohne Kontextoperator steigt die Anzahl der nachgewiesenen Dokumente um einen Faktor zwischen 1:4 und 1:5. Es ist plausibel, daß bei einem solchen Anstieg der Ballastanteil unzumutbar hoch steigt. Bei der Verwendung zweier Einfachdeskriptoren, verbunden mit dem Kontextoperator “gleicher Satz” steigt dagegen die Ballastquote nur noch um etwa 50 %. “Diese Quote zeigt klar die Vorteile komplexer Deskriptoren im Hinblick auf die *precision*, schließt ihren Ersatz durch den Kontextoperator “gleicher Satz” jedoch nicht mehr von vornherein aus, wenn die Aufwandsmessungen deutliche Unterschiede ergeben”[2].

6.3 Studie 1: Linguistische Beschreibung potentieller Fachtermini

Parallel zur Studie des frz. Korpus wurde der dt. Text desselben Ausschnitts aus dem EPÜ für eine linguistische Beschreibung potentieller Fachtermini herangezogen. Die-

1. Krause 1987, S. 17.
2. Krause 1987, S. 192.

ser Text ist aus den an anderer Stelle (siehe Abschnitt 4.4.3.1 Bewertung der Testreihen, Seite 56) dargelegten Gründen besonders gut für die Aufgabe geeignet.

6.3.1 Maximale Nominalphrasen

Die einzelnen Sätze und Überschriften wurden intellektuell geparst, d. h. in Verbal- und Nominalphrasen zerlegt. Die Verbalphrasen wurden nicht weiter analysiert. Die lokalisierten maximalen Nominalphrasen wurden unter folgenden Gesichtspunkten weiter untersucht:

- Welche Arten maximaler Nominalphrasen kommen vor?
- Aus welchen Teilphrasen bestehen die maximalen Nominalphrasen?
- In welcher Häufigkeitsverteilung kommen die maximalen Nominalphrasen und die Teilphrasen vor?
- Wie hoch ist der Anteil der Nominalkomposita?
- Welche Strukturen sind unter dem Aspekt einer kontrastiven dt.-frz. Untersuchung interessant?
- Welche Phrasentypen sind eventuell Fachtermini des Gebiets?

Im Vergleich mit dem frz. Text fällt auf, daß wider Erwarten ca. 50 % der maximalen Nominalphrasen nur aus einem Nomen und einem Determinator bestehen.

Maximale Nominalphrasen, die aus einem Nomen und einem Determinator bestehen

(1) die Haftung: la responsabilité

(2) die Frist: le délai

(3) die Rechtsstellung: le statut juridique

(4) die Rechtspersönlichkeit: la personnalité juridique

Maximale Nominalphrasen mit substantivischem Genitivattribut

(5) die Hoheitsgebiete der Vertragsstaaten: le territoire des Etats contractants

(6) die Zuständigkeit des Verwaltungsrats: la compétence du Conseil d’administration

(7) das Statut der Beamten: le statut des fonctionnaires

(8) die Bewilligung der Ausgaben: l’autorisation des dépenses

Maximale Nominalphrasen mit adjektivischem Attribut

(9) die allgemeinen Vorschriften: les dispositions générales

(10) das europäische Patent: le brevet européen

(11) die Europäische Patentorganisation: l'Organisation Européenne des brevets

(12) das technische Gutachten: l'avis technique

Maximale Nominalphrasen mit Präpositionalgruppe

(13) eine Streitigkeit nach Absatz 2: les litiges visés au paragraphe 2

(14) die Veröffentlichung von Mitteilungen an die Öffentlichkeit: la publication d'indications pour le public

(15) Vorschläge für eine Änderung dieses Übereinkommens: tout projet de modification de la présente convention

(16) die Organe im Verfahren: les instances chargées des procédures

Maximale Nominalphrasen mit Partizipialkonstruktion

(17) der von den Parteien geschlossene Vertrag: le contrat conclu entre les parties

(18) das vom Verwaltungsrat festgelegte Verfahren: la procédure fixée par le Conseil d'Administration

(19) die in der Ausführungsordnung vorgeschriebene Frist: le délai prévu par le règlement d'exécution

(20) das diesem Übereinkommen beigefügte Protokoll über Vorrechte und Immunitäten: le protocole sur les privilèges et immunités annexé à la présente convention

(21) die in Absatz 2 genannten Personen: les personnes visées au paragraphe 2

Maximale Nominalphrasen mit Adjektivphrase

(22) die zur Durchführung ihrer Aufgaben erforderlichen Vorrechte und Immunitäten: les privilèges et immunités nécessaires à l'accomplissement de leur mission

(23) das in der Bundesrepublik Deutschland zuständige Gericht: les juridictions compétentes de la République fédérale d'Allemagne

(24) alle für die Tätigkeit des Europäischen Patentamts zweckmäßigen Maßnahmen: toutes mesures utiles en vue d'assurer le fonctionnement de l'Office européen des brevets

(25) ein den Vertragsstaaten gemeinsames Recht für die Erteilung von Erfindungspatenten: un droit commun aux Etats contractants en matière de délivrance de brevets d'invention

Maximale Nominalphrasen mit Koordination

(26) bewegliches und unbewegliches Vermögen: des biens immobiliers et mobiliers

(27) natürliche oder juristische Personen: des personnes physiques et morales

(28) die verwaltungsmäßige und finanzielle Selbständigkeit: l'autonomie administrative et financière

(29) Entwürfe für allgemeine Durchführungsbestimmungen und Beschlüsse: tout projet de règlementation générale ou de décision

(30) Berichtigungs- und Nachtragshaushaltspläne: tout budjet modificatif ou additionnel

(31) Streitsachen zwischen der Organisation und den Bediensteten des Europäischen Patentamts: des litiges entre l'Organisation et les agents de l'Office Européen des brevets

6.3.2 Klassifikation der maximalen Nominalphrasen

6.3.2.1 Beschreibung

Die Beschreibung der maximalen Nominalphrasen erfolgt durch zwei Kriterien: die Phrasenlänge und das Bildungsmuster.

Bestimmung der Phrasenlänge

Die Länge einer Nominalphrase wird aus der Anzahl der Wörter der Hauptwortarten - Nomen, Adjektiv, Partizip - bestimmt. Nominalkomposita werden, obwohl sie aus mehreren Bestandteilen zusammengesetzt sind, als lexikalische Einheiten der Länge 1 behandelt, da sie keine Separatoren enthalten.

- Länge 1 (eingliedrig): Patent, Frist, Vertragsstaat, Beschwerdekammer
- Länge 2 (zweigliedrig): Große Beschwerdekammer, Mitglieder des Verwaltungsrats, Aufsicht über das Personal
- Länge 3 (dreigliedrig): Erlaß interner Verwaltungsvorschriften, Ernennung hoher Beamter, Eintragungen in das europäische Patentregister, außervertragliche Haftung der Organisation
- Länge 4 (viergliedrig): Europäisches Recht für die Erteilung von Patenten

Bestimmung der Bildungsmuster

Die Phrasenbildungsmuster geben die Art und die Reihenfolge der grammatischen Kategorien an, z. B.

- AN: Adjektiv Nomen: juristische Person

- Ngen[1]N: Nomen Det-Gen[2] Nomen: Deckung der Ausgaben

Aus dem Ausschnitt des EPÜ wurden 266 Nominalphrasen auf ihre Länge untersucht. Nominalphrasen, die eine Koordination enthalten, sind gesondert aufgeführt, da ihre Länge nicht eindeutig bestimmt werden kann.

Tabelle 26. Länge der maximalen Nominalphrasen in einem Ausschnitt des EPÜ-Korpus

Länge	**Anzahl (absolut)**	**in %**
Länge 1	114	42,86
Länge 2	68	25,56
Länge 3	38	14,29
Länge 4	16	6,015
Länge 5	5	1,88
Länge 6	4	1,50
Länge 7	1	0,38
Koordination	20	7,52
insgesamt	**266**	

6.3.2.2 Zerlegung

Die maximalen Nominalphrasen wurden in kleinere syntaktisch und semantisch zusammenhängende Teilphrasen zerlegt. Wie aus der Tabelle 1 ersichtlich ist, liegt der Hauptanteil der maximalen Nominalphrasen bei einer Länge zwischen 1 und 4, nur 11% sind länger. Es fällt auf, daß sich mehr als als drei Viertel dieser Phrasen in kleinere Teilphrasen (meist zweigliedrige) zerlegen lassen.

Da die eingliedrigen Nominalphrasen besonders häufig sind, wird dort zwischen einfachen Nomen und Komposita unterschieden, ein Aspekt, der bei der Untersuchung der Übersetzungsäquivalenzen besonders interessant ist.

Häufigste Bildungsmuster[3]

6.3.2.2.1 NP-Muster der Länge 1

N: einfaches Nomen

die Frist: le délai

1. Das Kürzel *gen* steht für zwei Dinge: 1. Determinator im Genitiv, 2. das darauffolgende Nomen steht im Genitiv.

2. Det-Gen: Determinator im Genitiv.

3. Wortklassennotation, sofern nicht explizit angegeben: N: Nomen; A: Adjektiv; A-Gen: Adjektiv im Genitiv; Praep: Präposition; Det: Determinator, Det-Gen: Determinator im Genitiv; N-Gen: Nomen im Genitiv. Fakultative Bestandteile stehen in Klammern, "/" steht für "ausschließliches Oder".

das Präsidium: le Bureau

Komp: Nominalkompositum

die Stimmenwägung: la pondération des voix

die Rechnungsprüfung: la vérification des comptes

6.3.2.2.2 NP-Muster der Länge 2

NPraepN: N Praep (Det) Nomen

die Eintragung in der Verfahrenssprache: l'inscription dans la langue de la procédure

die Teilnahme von Beobachtern: la participation d'observateurs

NgenN: Nomen Det-Gen Nomen

die Bewilligung der Ausgaben: l'autorisation de dépenses

die Ausführung des Haushaltsplans: l'éxécution du budget

AN: Adjektiv Nomen

die vorläufige Haushaltsführung: le budget provisoire

die territoriale Wirkung: la portée territoriale

6.3.2.2.3 NP-Muster der Länge 3

ANgenN: A N Det-Gen N-Gen

zerlegbar in: [A N] [Det-Gen N-Gen]

[eigene Mittel] [der Organisation]: ressources propres de l'Organisation

[die vertragliche Haftung] [der Organisation]: la responsabilité contractuelle de l'Organisation

NgenAN: N (Det-Gen) A-Gen N-Gen

zerlegbar in: [N] [Det-Gen A-Gen N-Gen] bzw. [N] [A-Gen N-Gen][1]

[die Dienststellen] [des Europäischen Patentamts]: les agences de l'Office Européen des brevets

[die Ernennung] [höherer Beamter]: la nomination du personnel supérieur

NPraepAN: N Praep (Det) A N

zerlegbar in: [N Praep[2]] [(Det) A N]

1. Zerlegung für N A-Gen N-Gen.

[Mittel für] [unvorhergesehene Ausgaben]: crédits pour dépenses imprévisibles

[die Eintragungen in] [das europäische Patentregister]: les inscriptions au Registre européen des brevets

NgenNgenN: N Det-Gen N-Gen Det-Gen N-Gen

zerlegbar in: [N Det-Gen N-Gen] [Det-Gen N-Gen] oder in: [N] [Det-Gen N-Gen Det-Gen N-Gen]

[die Unabhängigkeit] [der Mitglieder der Kammern]: l'indépendance des membres des chambres

NPraepNgenN: N Praep (Det) N Det-Gen N-Gen

zerlegbar in: [N Praep] [(Det) N Det-Gen N-Gen]

[Vorschläge für] [eine Änderung dieses Übereinkommens]: tout projet de modification de la présente convention

6.3.2.2.4 NP-Muster der Länge 4

ANPraepN(Praep/Gen)N: A N Praep N Praep/Det-Gen N/N-Gen

zerlegbar in: [A N Praep] [N Praep/Det-Gen N/N-Gen]

[das europäische Recht für] [die Erteilung von Patenten]: le droit européen de délivrance des brevets

NgenNgenAN: N Det-Gen N-Gen (Det-Gen) A-Gen N-Gen

zerlegbar in: $_1$[N]$_1$ $_2$[Det-Gen N (Det-Gen) $_3$[A-Gen N-Gen]$_3$]$_2$

[der Vorschlag] [des Präsidenten [des Europäischen Patentamts]]: la proposition du Président de l'Office européen des brevets

6.3.3 Fazit

Es ist damit zu rechnen, daß der größte Teil der potentiellen Fachtermini im dt. Teil des EPA-Korpus in Form von Nomen bzw. Nominalkomposita realisiert ist. Diese Annahme wird gestützt durch die Ergebnisse der Forschungen zur dt. Fachsprache.

Darüber hinaus werden die zweigliedrigen Phrasenmuster extrahiert, die im EPÜ-Ausschnitt häufig auftraten und in der Literatur zur dt. Fachterminologie als Bildungsmuster für Fachtermini angeführt werden. Außerdem sind kürzere Phrasen häufiger und eignen sich deshalb eher für eine Auswertung hinsichtlich der Frequenz.

2. Für die Zuordnung der Präposition gibt es zwei Möglichkeiten: zum darauffolgenden Nomen (da sie den Kasus regiert) oder zum vorhergehenden Nomen (da sie eine Subkategorisierung des vorhergegangen Nomens ist).

Die Extraktion maximaler Nominalphrasen erscheint im Kontext der vorliegenden Studie, die unter einem kontrastiven Aspekt steht, nicht sinnvoll. In der Analyse des frz. Teils des Korpus wurde kürzeren Komposita (d. h. mit weniger Gliedern) der Vorzug vor längeren gegeben. Die dt. Nominalkomposita haben sehr oft ein frz. Äquivalent, das aus mehreren Einheiten besteht. Eine Extraktion kürzerer Phrasen aus dem dt. Korpus gewährleistet die Parallelität zur Untersuchung des frz. Korpus.

6.4 Studie 2: Suche nach einem Werkzeug

Die maschinelle Extraktion der potentiellen Fachtermini erfordert ein Werkzeug, das zwei Aufgaben erfüllt:

- Lemmatisierung
- Extraktion

Für das frz. Korpus stand mit INTEX ein System zur Verfügung, das beide Teilaufgaben abdeckt. Für das dt. Korpus wurde eine andere Lösung gewählt. Es bestand die Möglichkeit, die dt. Texte mit einem sehr leistungsfähigen System (CISLEX-System) zu lemmatisieren. Das CISLEX-Sytem sieht aber keine Extraktion von Strukturen vor, so daß die zweite Teilaufgabe gesondert bearbeitet wurde.

6.4.1 Das CISLEX-System

Das am Centrum für Informations- und Sprachverarbeitung (CIS) der Universität München durchgeführte CISLEX-Projekt hatte zum Ziel, den Wortbestand des Deutschen in Form eines elektronischen Wörterbuchsystems zu erfassen. Das erstellte Lexikon (CISLEX) ist Grundlage für ein Lemmatisierungsprogramm, mit dem jede Wortform eines Textes (d. h. jede Buchstabenfolge, die zwischen Separatoren auftritt) identifiziert und auf seine Grundform zurückgeführt werden kann. Eine ausführliche Beschreibung des Lexikons und des Lemmatisierungsverfahrens findet sich in Maier-Meyer (1995).

Das CISLEX ist das umfassendste Lexikon des Deutschen. Auf der Basis verfügbarer Wortlisten wurde ein Grundstock von Lemmata angelegt, der durch den Abgleich mit gängigen Wörterbüchern ergänzt wurde. Der Wortbestand wird durch Korpusuntersuchungen ständig erweitert und aktualisiert. Als Korpusmaterial dienen aktuelle Texte (z. B. Tageszeitungen, Fachzeitschriften usw.) und literarische Texte, die in maschinenlesbarer Form zur Verfügung stehen.

6.4.1.1 Die Lexika

Im CISLEX werden auf der obersten Ebene vier Typen von Wortformen unterschieden, die in den entsprechenden Teillexika (Modulen) erfaßt sind.

1. Das **deutsche Kernlexikon** (CISLEX-DKL) enthält die einfachen und komplexen Wortformen[1] und ist im wesentlichen das deutsche Pendant zum frz. DELA.

2. Das **Namenslexikon** (CISLEX-EN) enthält Eigennamen aus den verschiedensten Bereichen (Vor- und Nachnamen, geographische Bezeichnungen und deren Ableitungen, Firmennamen usw.).

3. Das **Fremd- und Fachwörterbuch** (CISLEX-FF) enthält fremdsprachliche Wörter, die sich nicht in das deutsche morphologische Klassifikationsschema[2] integrieren lassen und Fachwörter, d. h. Wörter, die nicht zum allgemeinen dt. Wortschatz gehören, wie er in den großen Wörterbüchern (Duden, Wahrig, Mackensen usw.) erfaßt ist.

4. Das **Lexikon der Sonderformen** (CISLEX-SF) enthält Abkürzungen und Akronyme.

6.4.1.1.1 Einfache und komplexe Formen

Für die vorliegende Anwendung war in erster Linie das deutsche Kernlexikon und seine Unterscheidung in einfache und komplexe Formen relevant. Im Französischen ermöglicht das orthographische Kriterium der Getrennt- und Zusammenschreibung eine Unterscheidung zwischen einfachen und komplexen Formen. Im Deutschen ist es dagegen im allgemeinen nicht möglich, orthographisch die Komposita von den Simplizia zu unterscheiden. In den traditionnellen Grammatiken werden die Begriffe *einfach* (bzw. Simplex) und *komplex* meist mit semantischen Kriterien definiert. Da das DKL in erster Linie ein morphologisch orientiertes Lexikon des Deutschen ist, scheiden solche Kriterien von vornherein aus.

Sowohl bei der Derivation als auch bei der Komposition sind die morphologischen Eigenschaften eines Wortes durch dasjenige (Nicht-Flexions-) Morphem bestimmt, das am Ende des Wortes steht. Dies gilt auch für die Präfigierung. Diese Besonderheit der dt. Wortbildung ist die Grundlage für die Definition der einfachen und komplexen Formen im CISLEX.

> Ein Wort W ist eine **einfache Form** genau dann, wenn es keine sinnvolle Zerlegung $W = W_1 W_2$ gibt, so daß W_1 eine Folge von Morphemen ist und W_2 ein Wort mit denselben morphologischen Eigenschaften wie W.[3]

Umgekehrt ergibt sich als Definition für *komplexe Form*:

1. Definition der einfachen und komplexen Formen siehe Abschnitt 6.4.1.1.1 Einfache und komplexe Formen, Seite 127.

2. Beschrieben in: Maier-Meyer 1995, S. 45.

3. Definition in: Maier-Meyer 1995, S. 31.

> Ein morphologisch komplexes Wort W = W_1 ... W_n ist eine komplexe Form genau dann, wenn W_1 ... W_{n-1} Morpheme sind und W_n ein Wort mit denselben morphologischen Eigenschften wie W ist.[1]

Wie im frz. DELA-System wird auch im dt. Kernlexikon unterschieden zwischen einem Lexikon der einfachen Formen (DKL-EF), einem Lexikon der komplexen Formen (DKL-KF) und einem Lexikon der flektierten Formen (DKL-FLEX), das auf den beiden erstgenannten basiert.

6.4.1.1.2 Die Wortarten des CISLEX

Die Wortarten des deutschen Kernlexikons resultieren aus einer Mischklassifikation nach syntaktischen und morphologischen Kriterien, wobei die syntaktischen Kriterien als primär gegenüber den morphologischen betrachtet werden. Das hat zur Folge, daß insbesondere bei den Funktionswörtern zahlreiche Mehrfachklassifikationen auftreten, da in diesem Bereich viele Wörter an verschiedenen syntaktischen Positionen und mit unterschiedlichen syntaktischen Funktionen auftreten können. Auf diesen Punkt wird im weiteren Verlauf der Studie, insbesondere bei der Beschreibung der Verwendung des CISLEX für die Extraktion potentieller Fachtermini, noch genauer eingegangen werden.

Im CISLEX werden flektierende Wortarten (Nomen, Adjektiv, Determinator, Pronomen) und nicht-flektierende Wortarten (Adverb, Partikel, Präposition, Konjunktion, Interjektion und Verbpartikel) unterschieden. Das Grundformenlexikon DKL-EF enthält ca. 70 000 Einträge, von denen etwas mehr als die Hälfte Nomen sind. Aus der morphologischen Kodierung der DKL-EF-Formen wurden automatisch die dazugehörigen Vollfomen generiert (ca. 530 000 Einträge im FLEX-Lexikon).

6.4.1.1.3 Die Kodierung

Die Kodierung der Einträge berücksichtigt folgende Merkmalsklassen:

- **Typ**: Normale einfache Form, Kompositum, Eigenname, Sonderform, unbekannt.
- **Kategorie** (Wortklasse): z. B. Adjektiv, Nomen, Verb, Präposition usw.
- **morphologische Kodierung**: je nach Wortklasse Angaben zu Genus, Flexion usw.
- **syntaktische Kodierung**: z. B. Kasusforderungen von Präpositionen
- **semantische Kodierung**: Aufzählung semantischer Merkmale[2].

Wenn ein Lexem mehrere Merkmale einer Merkmalsklasse aufweist (z. B. im Falle einer Wortklassenambiguität oder einer morphologischen Ambiguität), so werden diese Merkmale im selben Eintrag kodiert.

---

1. Definition in: Maier-Meyer 1995, S. 31/32.
2. Vgl. dazu Langer 1996.

Für die vorliegende Untersuchung war neben den Angaben für Typ und Wortklasse vor allem die morphologische Kodierung[1] relevant. Je nach Wortart werden unterschiedliche morphosyntaktische Eigenschaften angegeben. So werden beispielsweise bei den Nomen Angaben zum Kasus, Numerus, Genus und Deklination gemacht, bei den Verben Angaben zu Person, Numerus, Tempus, Modus und Konjugation. Die einzelnen morphologischen Eigenschaften eines Lexems werden in einem Kode zusammengefaßt.

Beispiele:

Das Lexem mit der Form *Übereinkommen* hat zwei Lesarten, die sich in der morphologischen Kodierung unterscheiden. Es handelt sich um eine einfache Form (EF) der Wortklasse Nomen (N). Die unterschiedlichen morphologischen Kodierungen werden vorgenommen, weil es sich einmal um ein “gewöhnliches” Nomen (Kode 307), im zweiten Fall um die substantivierte Form des Verbs *übereinkommen* (Kode 305) handelt. In beiden Fällen wird auf die Form *Übereinkommen* lemmatisiert.

<form>Übereinkommen

typ=EF kat=N mor=305 syn=0 sem=[] >übereinkommen</lemma>

typ=EF kat=N mor=307 syn=0 sem=[] pref=1>übereinkommen</
lemma>

Der Kode 307 steht für:

Nominativ, Dativ und Akkusativ Singular Neutrum und Nominativ, Genitiv, Dativ und Akkusativ Plural Neutrum

Der Kode 305 steht für:

Nominativ, Dativ und Akkusativ Singular Neutrum.

Der Unterschied zwischen den beiden Nomen besteht darin, daß die vom Verb abgeleitete Form nicht in den Plural gesetzt werden kann.

Das Lexem *gemäß* ist eine einfache Form und kann eine Präposition (P) mit dem morphologischen Kode 0 oder ein Adjektiv (A) mit dem morphologischen Kode 341 sein. In beiden Fällen wird auf die Form *gemäß* lemmatisiert.

<form>gemäß

typ=EF kat=P mor=0>gemäß</lemma>

typ=EF kat=A mor=341>gemäß</lemma>

Der Kode 0 steht bei nichtflektierenden Wortklassen, der Kode 341 steht für den prädikativen Gebrauch (eines Adjektivs) im Positiv.

1. Vgl. dazu Maier-Meyer 1995, Kap. 2.3.

6.4.1.2 Lemmatisierung

Der Lemmatisierungsalgorithmus ist so ausgelegt, daß lediglich eine Segmentierung in Sätze in einem Vorverarbeitungsschritt (spezielles Satzendeerkennungsprogramm[1]) vor der eigentlichen Lemmatisierung durchgeführt werden muß. Der Algorithmus unterscheidet sich von herkömmlichen Verfahren vor allem unter zwei Aspekten:

- Zur Identifizierung von Wortformen wird ein **umfangreiches Lexikonsystem** verwendet, dadurch werden die Fehler beim Look-up auf das Auftreten lexikalischer Ambiguitäten beschränkt.
- **Sonder- und Mischformen** (numerische Ausdrücke, Ausdrücke mit Sonderzeichen etc.)[2], die von anderen Verfahren oft ignoriert werden, werden umfassend behandelt.

Der Lemmatisierungsalgorithmus weist jeder im Text auftretenden Wortform ihr Lemma bzw. bei Ambiguitäten ihre Lemmata zu. Das Lemma für einfache Formen ist der entsprechende Eintrag im FLEX-Lexikon und besteht aus Lemmaform, Lemmaklasse und morphosyntaktischen Merkmalen. Das Lemma für komplexe Formen enthält zusätzlich noch die entsprechende Segmentierung.

6.4.1.2.1 Kompositaanalyse

Bei der Segmentierung wird versucht, die Wortform in eine Struktur der Form [Vorderglied] [Vorderglied]*[3] [Hinterglied] zu zerlegen. Ein Hinterglied kann ein Eigenname oder eine einfache Form sein, die einer bestimmten Wortklasse (z. B. Nomen, Verb usw.) angehört. Der Segmentierungsalgorithmus versucht zuerst ein Suffix abzutrennen, das als mögliches Hinterglied in Frage kommt, was durch Look-up im FLEX-Lexikon überprüft wird. Wenn ein Suffix gefunden wurde, versucht der Algorithmus den verbleibenden Anfang entweder als Vorderglied oder als Folge von Vordergliedern zu identifizieren. Dieses Verfahren wird für alle möglichen Segmentierungen der Wortform wiederholt und liefert außer im Fall von Zahlausdrücken die gewünschten Lemmatisierungen. Zahlausdrücke werden durch ein anderes Verfahren lemmatisiert[4].

Grammatik für die Segmentierung[5]

```
komplexeForm -> Vorderglied + Hinterglied

komplexeForm -> ZahlForm

Vorderglied -> Fugenform|Präfix|Eigenname (Vorderglied)
```

1. Schicht 1994.
2. Vgl. dazu Maier-Meyer 1995, Kap.3.4, “|” steht für alternative, “(“ und “)” für fakultative Konstituenten.
3. Das Zeichen “*” bedeutet, daß das davorstehende Element n-mal wiederholt werden kann, mit n>=0.
4. Kap. 3.3.1.3.
5. Aus: Maier-Meyer 1995, S. 102.

```
Hinterglied -> Nomen|Verb|Adjektiv|Adverb|Suffix
```

6.4.1.2.2 Disambiguierung

Die Disambiguierung im CISLEX-System ist als regelbasiertes Verfahren implementiert. Die Menge der möglichen Analysen wird zuerst mit Hilfe linguistisch motivierter Regeln soweit wie möglich eingeschränkt. Bei den verbleibenden Mehrdeutigkeiten wird versucht, diese mit Hilfe von Präferenzregeln weiter einzuschränken. So ergibt sich insgesamt eine Hierarchie von Filtern, die nacheinander angewendet werden.

Arten der auftretenden Ambiguitäten und Heuristiken zur Disambiguierung sind von Mayer-Meyer[1] dargestellt. Als Disambiguierungsfilter werden genutzt:

- Orthographie: z. B. Groß/Klein-Filter
- Kontext: z. B. Kongruenzketten, Kontexte für bestimmte Wortarten
- Präferenzen: z. B. Präferenz für lexikalisierte Formen, Präferenz bei der Kompositasegmentierung für möglichst wenig Glieder

6.4.1.3 Die Extraktion von Strukturen

Das Extraktionsprogramm (in Perl) hat als Eingabe einen Text, der mit dem CISLEX-System lemmatisiert wurde und eine Liste von Strukturmustern, die in diesem Text lokalisiert werden sollen. Die Strukturmuster sind als Abfolgen von Kategorie(n) angegeben, ergänzt durch die Angabe des Typs (einfache oder komplexe Form).

Die Ausgabe des Programms ist eine Liste bestehend aus:

- Strukturen mit Lemmatisierung auf Grundform und Angabe zur Position im Text (Satznummer etc.)
- Angaben zur Ambiguität einer Struktur (bzgl. Wortklasse, morphologischer Kodierung, Segmentierung).

Das Progamm kann die morphologische Kodierung analysieren, z. B. um die Kongruenz zwischen Nomen und Adjektiv oder das Vorliegen eines bestimmten Kasus zu überprüfen.

6.4.2 Test

Ein Ausschnitt aus dem EPÜ-Korpus wurde mit dem CISLEX lemmatisiert und mit dem Extraktionsprogramm bearbeitet. Die Ausgabe wurde unter folgenden Gesichtspunkten analysiert:

Abdeckung des CISLEX-Wörterbuchs

1. Mayer-Meier 1995, Kap. 3.6.

Qualität der Lemmatisierung und der Segmentierung der Nominalkomposita

Korrektheit der extrahierten Strukturen.

Das Testkorpus ist ein ca. 7 000 Wörter umfassender Ausschnitt aus dem EPÜ (Art. 1-61). Die Satzendeerkennung aus dem CISLEX-System war, von wenigen spezifischen Abkürzungen abgesehen, ausreichend. Die Lemmatisierung erreichte eine Korrektheit von fast 100 %. Es wurden alle Wortformen erkannt (die Kategorie *UK* für unbekannt trat nicht auf). Von 7 000 Wörtern waren 1097 einfache Nomen und 604 Nominalkomposita.

Extrahierte Strukturmuster

1) N:einfaches Nomen

2) Komp: Nominalkompositum

3) AN:Adjektiv Nomen

5) NgenN: Nomen Det-Gen Nomen-Gen

4) NPraepN: N Praep (Det) Nomen

6.4.2.1 Mögliche Fehlerquellen

6.4.2.1.1 Ambiguität

Das Extraktionsprogramm unterscheidet bei der Ausgabe ambige und nicht-ambige Strukturen. Eine Struktur gilt als ambig, wenn mindestens ein Bestandteil mit mehr als einer Wortklasse oder mehr als einem morphologischen Kode versehen ist. Es wurden untersucht:

- die Arten der Ambiguitäten
- die Frequenz der Ambiguitäten
- die Auswirkung auf die Korrektheit der ausgegebenen Struktur
- Lösungsmöglichkeiten.

Die angegebenen Zahlen beziehen sich nicht auf die Gesamtheit der im Testkorpus vorkommenden Wortformen, sondern auf die extrahierten Strukturen.

1 Morphologische Ambiguität

Ein Wort gilt hier als morphologisch ambig, wenn es bei gleichbleibender Wortklasse mehr als einen morphologischen Kode aufweist. Es handelt sich meist um Fälle wie z. B. die Form *Übereinkommen*, die als Nomen oder als substantivierte Form des Verbs *übereinkommen* interpretiert werden kann. So wiesen beispielsweise von den einfa-

chen Nomen des Testkorpus ca. 8 % morphologische Ambiguitäten auf, die aber keinen Einfluß auf die korrekte Zuweisung der Wortklasse und der Grundform haben. Sie können deshalb für den weiteren Verlauf der Studie vernachlässigt werden.

2 Wortklassen-Ambiguität

Eine Wortklassenambiguität liegt dann vor, wenn ein Wort vom Lemmatisierungsprogramm mit meht als einer Wortklasse versehen wurde. Es wurden die Ambiguitäten von Nomen und Adjektiven untersucht, wobei in allen Fällen unter den angegebenen Wortklassen auch die richtige, d. h. die im jeweiligen Kontext realisierte, enthalten war.

Von 895 einfachen Nomen wurden 11 (1.2 %) mit mehr als einer Wortklasse ausgezeichnet. Im Testkorpus sind dies ausschließlich Wortformen, die am Satzanfang (bzw. am Anfang einer Überschrift) stehen. Da im Dt. der erste Buchstabe des ersten Wortes eines Satzes groß geschrieben wird, besteht an dieser Stelle immer eine Ambiguität: es kann sich um eine Wortform handeln, deren erster Buchstabe immer, d. h. nicht nur am Satzanfang, groß geschrieben wird oder um eine Wortform, deren erster Buchstabe normalerweise klein geschrieben wird.

Unter den komplexen Nomina (Nominalkomposita) des Testkorpus traten keine Wortklassenambiguitäten auf.

Von den im Testkorpus auftretenden Adjektiven wurden nur diejenigen untersucht, die in der Struktur “Adjektiv Nomen” auftreten. Von den 340 Adjektiven waren 24 (darunter 17 verschiedene Grundformen), d. h. 7 % bzw. 5 % ambig. Der größte Teil davon sind Adjektive, die auch eine Partizipialform sein können und deswegen auch mit der Wortklasse Verb ausgezeichnet sind. Daneben traten zwei Fälle von Ambiguitäten zwischen Präposition und Adjektiv (*vorbehaltlich*, *unter*) auf, die durch den Kontext (Kongruenz) aufgelöst werden können.

3 Segmentierungsambiguität

Eine Ambiguität liegt dann vor, wenn der beschriebene Segmentierungsalgorithmus (siehe Abschnitt 6.4.1.1.1 Einfache und komplexe Formen, Seite 127) mehr als eine Zerlegung erzeugt.

Von 569 Nominalkomposita wiesen 35 (darunter 18 verschiedenen Grundformen) (6 bzw. 3.1 %) eine Segmentierungsambiguität auf, die eine morphologische Ambiguität zur Folge hat. In 31 Fällen handelt es sich um Komposita im Plural, bei denen als Grundwort sowohl der Singular als auch der Plural aufgeführt wird, z. B.:

Dienststellen -> Dienst+stelle bzw. Dienst+stellen

Diagnostizierverfahren -> Diagnostizier+Verfahren (Sing.) bzw. Diagnostizier +Verfahren (Plur.)

In den restlichen vier Fällen ist das Grundwort ein von einem Präfixverb abgeleitetes Substantiv, z. B.:

Aufrechterhaltung -> Aufrechter+Haltung bzw. Aufrecht+Erhaltung

Segmentierungsambiguitäten können im weiteren Verlauf der Studie vernachlässigt werden, da sie selten sind und der Anteil falscher Segmentierungen sehr gering ist (0.06 % aller Nominalkomposita des Testkorpus).

6.4.2.1.2 Mehrwortverbindungen adverbialer oder präpositionaler Funktion

In den Lexika des CISLEX-Systems werden einfache und komplexe Formen unterschieden. Im Gegensatz zum frz. DELA-System sind im CISLEX in der jetzigen Version keine Mehrwortverbindungen erfaßt. Die Untersuchung des Testkorpus zeigt, daß solche Formen im Dt. auch vorkommen und vor allem im Zusammenhang mit der **maschinellen** Analyse eines Textes von Bedeutung sind. Für eine korrekte Bestimmung syntaktischer Bildungsmuster in einem Text ist die vorherige Identifizierung von inhaltlich und funktional zusammengehörigen Einheiten, die aus mehreren, getrennt geschriebenen Wörtern bestehen, notwendig, um die Extraktion von inkorrekten oder inhaltlich wenig aussagekräftigen Strukturen zu verhindern. Sucht man beispielsweise nach dem Strukturmuster “N Det-Gen N-Gen”, erhält man “Gebiet des Patentwesens” (vorkommend in: “auf dem Gebiet des Parentwesens”) , eine Phrase, die zwar korrekt ist, die aber als potentieller Fachterminus kaum mehr Information bietet als das Nomen (Patentwesen) allein.

Folgende Arten von Mehrwortverbindungen kamen vor:

- Präpositionen: z. B. auf der Grundlage von, in Bezug auf, im Verhältnis zu, im Sinne von
- Adverbien: z. B. von Amts wegen
- Eigennamen: z. B. Bundesrepublik Deutschland, Vereinigtes Königreich
- Teile von Funktionsverbgefügen: der Meinung sein, zur Schau stellen.

Aus dem Textkorpus wird ersichtlich, daß solche Einheiten sowohl allgemeinsprachlicher als auch fachsprachlicher Natur sein können. Zum Zwecke einer möglichst fehlerfreien automatischen Textanalyse sollten sie im Vorfeld der lexikalischen Analyse systematisch erfaßt werden. Dies wurde für Ausschnitte aus dem EPÜ- und aus dem EPA-Korpus durchgeführt.

6.4.2.1.3 Zuordnung einer falschen Wortklasse

Hier handelt es sich um Fälle, in denen einer Wortform nur eine Wortklasse zugeordnet wurde und diese Wortklasse nicht der im Text realisierten Wortklasse entspricht. Falsche Wortklassen traten nur bei Nomen auf und dort nur in zwei Fällen. Es handelt

sich um Eigennamen, die Institutionen bezeichnen und deren Bestandteile deshalb groß geschrieben werden:

E*uropäische* vorkommend in: das *Europäische Patentamt*, die *Europäische Patentorganisation* (51 Vorkommen im Testkorpus)

G*roße* vorkommend in: die G*roße Beschwerdekammer* (10 Vorkommen im Testkorpus).

6.4.2.1.4 Zerlegungsfehler

Die Extraktion der Phrasenmuster erfolgt durch einfaches Pattern-Matching ohne vollständige Syntaxanalyse. Das hat zur Folge, daß ein Teil der extrahierten Strukturen, die diesen Phrasenmustern entsprechen, keine korrekten Nominalphrasen sind, sondern Teilphrasen einer längeren Nominalphrase, die falsch zerlegt worden war. Dazu zählen auch Nominalphrasen, die eine Koordination enthalten. Diese Fehler können nur mit einer komplexeren syntaktischen Analyse, wie sie z. B. in den Systemen CTX oder DETECT (siehe Abschnitt 6.2.4 Vergleichende Evaluierung: DETECT vs CTX, Seite 117) existiert, behoben werden.

Beispiele:

erteilte Patente (Muster AN), vorkommend in: die nach diesem Übereinkommen erteilten Patente (les brevets délivrés en vertu de cette convention)

Züchtung von Pflanzen (Muster NPraepN, mit Koordination), vorkommend in: Verfahren zur Züchtung von Pflanzen oder Tieren (procédés d'obtention des végétaux ou d'animaux)

Ausgleich des Haushalts (Muster N Det-Gen N) vorkommend in: Ausgleich des Haushalts der Organsation (équilibre du budget de l'Organisation)

6.5 Studie 3: Anwendung des CISLEX-Systems

Dieser Untersuchung liegt der dt. Teil des Korpus zugrunde, dessen frz. Teil im vorhergehenden Kapitel für eine Extraktion der frz. Strukturen verwendet wurde. Dieses Teilkorpus wurde mit dem CISLEX lemmatisiert und auf das Vorkommen der dargestellten Phrasenbildungsmuster untersucht (Perl-Programm). Alle extrahierten Strukturen wurden auf ihre Korrektheit überprüft.

6.5.1 Ergebnisse

Das Lemmatisierungsprogramm erkannte den größten Teil der Wortformen, nur 1,75% der Wortformen wurden mit der Kategorie *unbekannt* ausgezeichnet. Hier handelt es sich in erster Linie um Fachwörter aus der Chemie (z. B. *Critobalit*, *Diastereomere*, *calciniert*). Als weitere unbekannte Wortformen kamen Eigennamen (z. B. *Ciba Geigy*) und Abkürzungen (z. B. *EPÜ*) vor.

Das CISLEX-System erkennt Bindestrichwörter[1] richtig. Koordinationsbindestriche wurden korrekt analysiert und die Bezugsformen richtig rekonstruiert (z. B. Verfahrens- und Sachansprüche -> Verfahrensansprüche und Sachansprüche).

Die Ergebnisse werden für jedes der sechs untersuchten Dokumente gesondert aufgeführt. Bei den Vorkommen eines Phrasenbildungstyps werden nicht ambige und ambige Formen unterschieden. Als *ambig* gelten Formen, die eine morphologische Ambiguität, eine Wortklassenambiguität oder eine Segmentierungsambiguität aufweisen (siehe Abschnitt 6.4.2.1.1 Ambiguität, Seite 132).

Da die Fehler, die bei der Analyse von einfachen und komplexen Nomina auftreten, anderer Art sind als die Fehler, die bei der Analyse von Mehrwortphrasen zu beobachten sind, werden die Ergebnisse getrennt für Nomina und Mehrwortphrasen aufgeführt.

6.5.1.1 Nomina

Für jedes Dokument werden sowohl für die einfachen Nomina als auch für die Nominalkomposita angegeben:

- die Gesamtzahl der Vorkommen (die in der Tabelle mit *gesamt* bezeichnete Spalte).
- die Anzahl der korrekt extrahierten Vorkommen (die in der Tabelle mit *korrekt* bezeichnete Spalte).
- die Anzahl der fehlerhaft extrahierten Vorkommen (die in der Tabelle mit *falsch* bezeichnete Spalte), z. B. Nomen, die fälschlicherweise als Nomen lemmatisiert wurden.

Die Zeile *insgesamt* gibt die Summe einer Spalte an.

Tabelle 27. Nomina im Dokument 841

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	507	500	7
einfaches Nomen (ambig)	56	46	10
Nominalkompositum (nicht ambig)	255	255	0
Nominalkompositum (ambig)	29	28	1
insgesamt	**847**	**829**	**18**

1. Das Verfahren ist in Maier-Meyer 1995, S. 116ff beschrieben.

Tabelle 28. Nomina im Dokument 1089

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	719	703	16
einfaches Nomen (ambig)	86	73	13
Nominalkompositum (nicht ambig)	242	241	1
Nominalkompositum (ambig)	62	62	0
insgesamt	**1109**	**1079**	**30**

Tabelle 29. Nomina im Dokument 2016

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	720	711	9
einfaches Nomen (ambig)	59	54	5
Nominalkompositum (nicht ambig)	218	215	3
Nominalkompositum (ambig)	58	58	0
insgesamt	**1055**	**1038**	**17**

Tabelle 30. Nomina im Dokument 2461

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	726	718	8
einfaches Nomen (ambig)	66	66	0
Nominalkompositum (nicht ambig)	262	262	0
Nominalkompositum (ambig)	25	25	0
insgesamt	**1079**	**1071**	**8**

Tabelle 31. Nomina im Dokument 2651

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	1089	1066	23
einfaches Nomen (ambig)	104	88	16
Nominalkompositum (nicht ambig)	405	403	2
Nominalkompositum (ambig)	18	18	0
insgesamt	**1616**	**1575**	**41**

Tabelle 32. Nomina im Dokument 2840

Strukturmuster	gesamt	korrekt	falsch
einfaches Nomen (nicht ambig)	813	800	13
einfaches Nomen (ambig)	98	86	12
Nominalkompositum (nicht ambig)	323	323	0
Nominalkompositum (ambig)	35	35	0
insgesamt	**1269**	**1244**	**25**

6.5.1.2 Mehrwortphrasen

Für jedes Dokument und für jeden Phrasenbildungstyp werden angegeben:

- die Gesamtzahl der Vorkommen eines bestimmten Typs (die in der Tabelle mit *gesamt* bezeichnete Spalte).
- die Anzahl der korrekt extrahierten Vorkommen (die in der Tabelle mit *korrekt* bezeichnete Spalte).
- die Anzahl der Vorkommen, die eine adverbiale oder präpositionale Mehrwortverbindung enthalten[1] (die in der Tabelle mit *enthält MWV* (für: Mehrwortverbindung) bezeichnete Spalte).
- die Anzahl der Fälle, in denen die vorkommende Struktur inkorrekt extrahiert wurde, da sie Teil einer längeren Phrase ist[2] (die in der Tabelle mit *Zerlegungsfehler* bezeichnete Spalte).
- die Anzahl sonstiger Fehler (die in der Tabelle mit *andere Fehler* bezeichnete Spalte), z. B. eine falsche Wortklasse[3].

Die Zeile *insgesamt* gibt die Summe einer Spalte an.

1. Vgl. Abschnitt 6.4.2.1.2 Mehrwortverbindungen adverbialer oder präpositionaler Funktion, Seite 134.
2. Vgl. Abschnitt 6.4.2.1.4 Zerlegungsfehler, Seite 135.
3. Vgl. Abschnitt 6.4.2.1.3 Zuordnung einer falschen Wortklasse, Seite 134.

Tabelle 33. Mehrwortphrasen im Dokument 841

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	177	121	3	53	0
A N (ambig)	70	52	0	17	1
NgenN (nicht ambig)	20	7	2	11	0
NgenN (ambig)	94	74	5	12	3
N P N (nicht ambig)	15	8	0	7	0
N P N (ambig)	14	5	1	8	0
N P Det N (nicht ambig)	0	0	0	0	0
N P Det N (ambig)	33	4	2	26	1
insgesamt	**423**	**271**	**13**	**134**	**5**

Tabelle 34. Mehrwortphrasen im Dokument 1087

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	184	168	0	16	0
A N (ambig)	81	70	1	10	0
NgenN (nicht ambig)	21	2	6	13	0
NgenN (ambig)	119	83	9	23	4
N P N (nicht ambig)	50	13	8	29	0
N P N (ambig)	20	5	7	8	0
N P Det N (nicht ambig)	1	0	0	1	0
N P Det N (ambig)	46	11	6	29	0
insgesamt	**522**	**352**	**37**	**129**	**4**

Tabelle 35. Mehrwortphrasen im Dokument 2016

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	191	177	0	14	0
A N (ambig)	80	69	0	11	0
NgenN (nicht ambig)	32	2	6	24	0
NgenN (ambig)	100	63	20	17	0
N P N (nicht ambig)	57	26	15	15	1
N P N (ambig)	30	10	11	9	0
N P Det N (nicht ambig)	0	0	0	0	0
N P Det N (ambig)	40	4	9	27	0
insgesamt	**530**	**351**	**61**	**117**	**1**

Tabelle 36. Mehrwortphrasen im Dokument 2461

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	283	248	1	34	0
A N (ambig)	91	74	1	16	0
NgenN (nicht ambig)	12	1	6	5	0
NgenN (ambig)	111	70	16	24	1
N P N (nicht ambig)	24	3	6	15	0
N P N (ambig)	8	1	3	4	0
N P Det N (nicht ambig)	0	0	0	0	0
N P Det N (ambig)	23	5	6	12	0
insgesamt	**552**	**402**	**39**	**110**	**1**

Tabelle 37. Mehrwortphrasen im Dokument 2651

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	201	163	0	38	0
A N (ambig)	76	55	4	17	0
NgenN (nicht ambig)	32	0	13	19	0
NgenN (ambig)	231	161	30	28	12
N P N (nicht ambig)	70	4	5	61	0
N P N (ambig)	11	1	5	5	0
N P Det N (nicht ambig)	0	0	0	0	0
N P Det N (ambig)	80	17	5	58	0
insgesamt	**701**	**401**	**62**	**226**	**12**

Tabelle 38. Mehrwortphrasen im Dokument 2840

Strukturmuster	gesamt	korrekt	enthält MWV	Zerlegungs-fehler	andere Fehler
A N (nicht ambig)	178	151	1	26	0
A N (ambig)	79	59	0	20	0
NgenN (nicht ambig)	14	2	3	9	0
NgenN (ambig)	137	111	8	15	3
N P N (nicht ambig)	32	13	4	15	0
N P N (ambig)	22	4	8	10	0
N P Det N (nicht ambig)	0	0	0	0	0
N P Det N (ambig)	63	10	12	41	0
insgesamt	**525**	**350**	**36**	**136**	**3**

6.5.1.3 Gesamtstatistik

In der Gesamtstatistik werden die Einzelergebnisse zusammengefaßt. Für jedes Strukturmuster und jedes Dokument ist die Anzahl der korrekten Vorkommen aufgeführt.

Tabelle 39. Gesamtstatistik

Strukturmuster	Summe aus 6 Dok.	Dok. 841	Dok. 1087	Dok. 2016	Dok. 2461	Dok. 2651	Dok. 2840
einfaches Nomen (nicht ambig)	4498	500	703	711	718	1066	800
einfaches Nomen (ambig)	413	46	73	54	66	88	86
Nominalkompositum (nicht ambig)	1699	255	241	215	262	403	323
Nominalkompositum (ambig)	226	28	62	58	25	18	35
A N (nicht ambig)	1028	121	168	177	248	163	151
A N (ambig)	379	52	70	69	74	55	59
NgenN (nicht ambig)	14	7	2	2	1	0	2
NgenN (ambig)	562	74	83	63	70	161	111
N P N (nicht ambig)	67	8	13	26	3	4	13
N P N (ambig)	26	5	5	10	1	1	4
N P Det N (nicht ambig)	0	0	0	0	0	0	0
N P Det N (ambig)	51	4	11	4	5	17	10
insgesamt	**8963**	**1100**	**1431**	**1389**	**1473**	**1976**	**1594**

6.5.2 Fazit

Das CISLEX-System erwies sich durch den Umfang des Lexikons und die Qualität der lexikalischen Analyse als effizientes Werkzeug für die Lemmatisierung des untersuchten Korpus. Die auftretenden Ambiguitäten sind in erster Linie morphologischer Natur. Sie sind die Folge einer sehr differenzierten Kodierung, die für den vorliegenden Zweck in den meisten Fällen nicht nötig war. Dies wird aus der Tatsache ersichtlich, daß der Korrektheitsgrad der ambigen und der nicht-ambigen Strukturen nur geringe Unterschiede aufweist.

Die Extraktion der Phrasen liefert für die Nomina und die Adjektiv-Nomen-Verbindungen zufriedenstellende Ergebnisse. Die anderen Arten von Mehrwortphrasen können auf diese Art zwar aus einem Text extrahiert werden, müssen aber in jedem Fall intellektuell überprüft werden. Dieser Ansatz ist jedoch bei einem größeren Korpus nicht praktikabel. Abhilfe kann hier nur durch Methoden geschaffen werden, in denen zumindest eine partielle syntaktische Analyse stattfindet. Die geringe Zahl der korrekten, d. h. der tatsächlich auftretenden Mehrwortphrasen (dies gilt nicht für die Nomen-Adjektiv-Verbindungen) mag verwundern. Sie bestätigt aber die Ergebnisse linguistischer Untersuchungen von Patenttexten: der Anteil sehr langer und komplexer Nomi-

nalphrasen ist dort bei weitem höher als in allgemeinsprachlichen Texten, d. h. die extrahierten Mehrwortphrasen waren in den meisten Fällen Teilphrasen einer längeren Nominalphrase. Dies kommt auch in dem hohen Anteil der Zerlegungsfehler zum Ausdruck.

Im untersuchen Korpus zeigt sich die Tendenz, daß bei der Bildung von Benennungen die in der Forschung zur deutschen Fachsprache erwähnten Mehrwortverbindungen im Vergleich mit den einfachen und komplexen Nomina eine untergeordnete Rolle spielen. Es müßte geprüft werden, ob diese Feststellung durch die Untersuchung eines größeren Korpus bestätigt werden kann.

7 Kontrastive Auswertung und mögliche Anwendungen

Vor der eigentlichen kontrastiven Auswertung werden zunächst die Ergebnisse der Extraktion aus dem frz. und dem dt. Korpus zusammengefaßt.

Die Arbeiten zur bilingualen Extraktion von Fachtermini, die vorgestellt wurden (siehe Abschnitt 2.2 Computerlinguistische Untersuchungen von Fachsprachen, Seite 10), basieren auf zwei Hypothesen: Es wird angenommen, daß die Fachtermini in den beiden Sprachen nach einer begrenzten Anzahl von Bildungsmustern aufgebaut sind und daß zwischen den extrahierten Termini Übersetzungsäquivalenzen vorliegen. Diese Hypothesen werden an Hand des EPA-Korpus überprüft. Darüber hinaus soll gezeigt werden, in welcher Form Ergebnisse einer kontrastiven Untersuchung in der Praxis genutzt werden können.

7.1 Gesamtauswertung der Extraktion

In der folgenden Tabelle sind die Ergebnisse der Extraktion für jede Sprache zusammengefaßt.

Dabei ist zu beachten:

- Die Extraktion aus dem frz. und aus dem dt. Korpus wurde mit unterschiedlichen Methoden durchgeführt. In beiden Fällen ist ein Teil der extrahierten Strukturen nicht korrekt (vgl. dazu für das frz. Korpus Abschnitt 5.5.2 Ergebnisse, Seite 99 bzw. für das dt. Korpus Abschnitt 6.5.1 Ergebnisse, Seite 135).
- Aus dem frz. Korpus konnten selten alle Vorkommen eines gegebenen Bildungsmusters extrahiert werden, da bei der Anwendung von INTEX die Linkskontexte und Rechtskontexte der Bildungsmuster aus den angegebenen Gründen eingeschränkt worden waren (vgl. Abschnitt 5.4.4 Probleme und Lösungen: Eignung für die Aufgabe und Anpassung, Seite 89).

  Für das frz. Korpus wird deshalb dieAnzahl der extrahierten Vorkommen angegeben, die **korrekt** sind sowie die Anzahl der **ergänzten** Strukturen (vgl. Abschnitt 5.5.2.2 Ergänzte Strukturen, Seite 102) angegeben.
- Das dt. Korpus wurde mit dem CISLEX-System lemmatisiert, die Extraktion wurde mit Hilfe eines Perl-Programms (siehe Abschnitt 6.4.1.3 Die Extraktion von Strukturen, Seite 131) durchgeführt.

  Für das dt. Korpus wird dieAnzahl der extrahierten Vorkommen angegeben, die **korrekt** sind.

Tabelle 40. Gesamtauswertung der Extraktion (frz. und dt. Korpus)

Strukturmuster	Summe aus 6 Dok.	Dok. 843/ 841	Dok. 1089/ 1087	Dok. 2018/ 2016	Dok. 2463/ 2461	Dok. 2653/ 2651	Dok. 2842/ 2840
französisch							
NA	1065	181	175	252	199	125	133
NdeN	1135	188	214	150	150	305	128
NàN	19	14	0	2	0	1	2
NPrepN	74	4	31	13	6	16	4
NN	69	36	1	3	17	2	10
NdeNA	180	29	23	51	41	20	16
NdeNdeN	112	8	27	19	11	30	17
NAA	48	11	10	16	5	4	2
NAdeN	74	9	14	5	14	27	5
NdeNdeNN/A	19	5	4	4	3	2	1
Sonstige	9	1	7	0	0	0	1
insgesamt	**2804**	**486**	**506**	**515**	**446**	**532**	**319**
deutsch							
einfaches Nomen (nicht ambig)	4498	500	703	711	718	1066	800
einfaches Nomen (ambig)	413	46	73	54	66	88	86
Nominalkompositum (nicht ambig)	1699	255	241	215	262	403	323
Nominalkompositum (ambig)	226	28	62	58	25	18	35
A N (nicht ambig)	1028	121	168	177	248	163	151
A N (ambig)	379	52	70	69	74	55	59
NgenN (nicht ambig)	14	7	2	2	1	0	2
NgenN (ambig)	562	74	83	63	70	161	111
N P N (nicht ambig)	67	8	13	26	3	4	13
N P N (ambig)	26	5	5	10	1	1	4
N P Det N (nicht ambig)	0	0	0	0	0	0	0
N P Det N (nicht ambig)	0	0	0	0	0	0	0
N P Det N (ambig)	51	4	11	4	5	17	10
insgesamt	**8963**	**1100**	**1431**	**1389**	**1473**	**1976**	**1594**

7.2 Was ist eine Übersetzungsäquivalenz?

Die Voraussetzung für die Zusammenführung einer Benennung in einer Sprache und ihres Gegenstücks in einer anderen Sprache ist ihre weitgehende inhaltliche Übereinstimmung, d. h. ihre Äquivalenz.

In der Übersetzungswissenschaft wird die Äquivalenz von Fachtermini folgendermaßen definiert:

> Zwei (durch Termini repräsentierte) Begriffe sind grundsätzlich dann als äquivalent zu betrachten, wenn sie in sämtlichen Begriffsmerkmalen übereinstimmen, d. h. wenn Identität der Begriffe vorliegt. Wenn zwei Begriffe in den wesentlichen Merkmalen übereinstimmen, aber in unwesentlichen voneinander abweichen, so besteht zwar keine Identität, dennoch lassen sich die Benennungen aufgrund der hochgradigen begrifflichen Entsprechung zusammenführen.[1]

Im Gegensatz dazu werden in der vorliegenden Arbeit Übersetzungsäquivalenzen aus parallelen Texten erhoben, d. h. aus einem Text in der Sprache $L_{1,}$ (bezeichnet mit $Text1_{L1}$) und der Übersetzung dieses Textes in der Sprache L_2 ($Text1_{L2}$). Dann gilt:

A_{L1}(Text1, Satz x) ist eine einfache oder komplexe Form der Sprache L_1, die in Text1 in Satz x vorkommt und

B_{L2}(Text1, Satz y) ist eine einfache oder komplexe Form B der Sprache L_2, die in Text1 in Satz y vorkommt.

Zwischen A_{L1} und B_{L2} liegt dann eine **Übersetzungsäquivalenz** vor, wenn A_{L1} durch B_{L2} übersetzt wurde. Es handelt sich also um Äquivalenzen, die vorerst nur in dem angegebenen Kontext Gültigkeit haben. Es ist zu überprüfen, inwieweit solche kontextuellen Äquivalenzen verallgemeinert werden können.

Die Übersetzungsäquivalenzen der Grundform (Lemma) einer einfachen oder komplexen Form werden für jedes Vorkommen dieses Lemmas bestimmt.

7.2.1 Ermittlung der Äquivalenzen

Die Äquivalenzen werden aus den satzweise alignierten Texten ($Text1_{L1}$ und $Text1_{L2}$) ermittelt. Ausgangspunkt dieser Untersuchung ist eine Liste der aus $Text1_{L1}$extrahierten Formen, der Sätze aus $Text1_{L1}$, in denen diese Formen vorkommen sowie der Sätze aus $Text1_{L2}$, mit denen die entsprechenden Sätze aus $Text1_{L1}$ aligniert sind. Daraus wird für jede Form A_{L1} der Ausgangsliste ermittelt, wie sie übersetzt wurde.

Es wurde festgestellt, daß in einem Text nicht immer alle Vorkommen einer Form A_{L1} durch *ein* zielsprachliches Äquivalent aus L_2 in einer1:1-Entsprechung wiedergegeben werden. Es kann jedoch davon ausgegangen werden, daß sich unter den zielsprachli-

1. Arntz & Picht 1982, S. 141.

chen Äquivalenten aus L_2 auch die Form befindet, die als *kanonische Übersetzung* gelten kann, zumal es sich bei den untersuchten Formen um potentielle Fachtermini handelt. Die kanonische Übersetzung ist in den meisten Fällen auch die häufigste Form.

Zur Veranschaulichung werden nun Übersetzungsbeispiele für Nominalphrasen gezeigt, die mit einer hohen Frequenz in den Texten auftraten und die eindeutig als Fachtermini einzuordnen sind[1]. Die Beispiele sind nach den Sprachrichtungen angeordnet, unabhängig davon, welche Sprache im einzelnen Fall die Quell- bzw. Zielsprache war. Die Struktur der untersuchten Nominalphrasen wird auch angegeben.

7.2.1.1 Beispiele für die Sprachrichtung Französisch-Deutsch

7.2.1.1.1 Übersetzung von "décision de révocation"

(1) décision de révocation: Widerrufsentscheidung (NdeN: Komp)

> Si l'on retient cette interprétation, il n'est pas nécessaire d'examiner si la prise d'une **décision de révocation** exclut l'octroi de la restitutio in integrum au titre de l'article 122 CBE.
>
> Angesichts dieser Auslegung bedarf es keiner Entscheidung darüber, ob durch eine **Widerrufsentscheidung** die Wiedereinsetzung in den vorigen Stand nach Artikel 122 EPÜ ausgeschlossen wird.

(2) décision de révocation: Entscheidung (NdeN: N)

> Une **décision de révocation** n'aurait donc en elle-même aucun sens; qui plus est, elle risquerait de donner lieu à un recours absurde dans lequel aussi bien les parties concernées que l'OEB perdraient leur temps et leur argent.
>
> Somit ist eine **Entscheidung** nicht nur an sich zwecklos; sie kann auch eine ebenso zwecklose Beschwerde nach sich ziehen, was wiederum mit einer Zeit- und Geldverschwendung für die Beteiligten und das EPA verbunden ist.

(3) décision de révocation: Entscheidung über den Widerruf (NdeN: NPDetN)

> Dans ces conditions, le contenu de ces documents ne doit pas être interprété comme constituant une **décision de révocation** au sens de l'article 106 (1) CBE.
>
> Unter diesen Umständen sollten sie von ihrer Substanz her nicht als **Entscheidung über den Widerruf** des Patents im Sinne des Artikels 106 (1) EPÜ verstanden werden.

1. Dies wurde durch Rücksprache mit einem Experten auf dem jeweiligen Fachgebiet bestätigt.

7.2.1.1.2 Übersetzung von “activité inventive”

(4) activité inventive: erfinderische Tätigkeit (NA: AN)

La division d'opposition a toutefois fait observer dans sa décision qu'elle considérait que l'invention revendiquée impliquait une **activité inventive**.

Eine **erfinderische Tätigkeit** sah die Einspruchsabteilung in ihrer Entscheidung bei der beanspruchten Erfindung aber gegeben.

(5) activité inventive: erfinderisch (NA: A)

Selon lui, l'objet des revendications de composition 4 et 5 de l'Annexe 1 était nouveau et impliquait une **activité inventive** par rapport au document (1), de même que celui des revendications 6 et 7, étant donné les “disclaimers” que comportent ces revendications et qui sont relatifs aux compositions spécifiques contenant du chlorure de lanthane.

Die Sachansprüche 4 und 5 in Anlage 1 seien gegenüber dem Dokument 1 sowohl neu als auch **erfinderisch**; dasselbe gelte für die Ansprüche 6 und 7, nachdem Disclaimer für die im Dokument 1 beschriebenen spezifischen Zusammensetzungen mit Lanthanchlorid aufgenommen worden seien.

7.2.1.1.3 Übersetzung von “maintien du brevet”

(6) maintien du brevet: Aufrechterhaltung des Patents (NdeN: NgenN)

... alors qu'au point 8, par exemple, il est prévu expressément que les agents des formalités peuvent se voir confier" la décision portant **maintien du brevet** européen tel qu'il a été modifié, conformément à l'article 102, paragraphe 3 de la CBE".

... vgl. hierzu auch die Nummer 8 der Mitteilung, in der die Formalsachbearbeiter ausdrücklich mit dem "Erlaß der Entscheidung über die **Aufrechterhaltung des** europäischen **Patents** in geändertem Umfang nach Artikel 102 (3) EPÜ" betraut werden.

(7) maintien du brevet: das Patent ... aufrechtzuerhalten (NdeN: Infinitivphrase)

A titre subsidiaire, elle demande le **maintien du brevet** avec une nouvelle limitation, à savoir la suppression des revendications 4 et 5.

Hilfsweise beantragt sie, das **Patent** mit der weiteren Einschränkung **aufrechtzuerhalten**, daß Ansprüche 4 und 5 gestrichen werden.

(8) maintien du brevet: Aufrechterhaltung (NdeN: Komp)

Ce n'est que dans les cas où les parties ne peuvent raisonnablement se prononcer de manière définitive dès le stade de la procédure orale sur la question du **maintien du brevet** dans sa forme modifiée, du fait par exemple que les modifications apportées sont si nombreuses ou si importantes qu'elles ne peuvent en apprécier immédiate-

ment toute la portée, qu'il serait nécessaire, même après la tenue de cette procédure orale, d'adresser aux parties la notification visée par la règle 58 (4) de la CBE, afin de leur donner des possibilités suffisantes de se faire entendre.

Nur in den Fällen, in denen es den Parteien nicht zuzumuten ist, eine endgültige Stellungnahme zur **Aufrechterhaltung** in geändertem Umfang bereits in der mündlichen Verhandlung abzugeben, etwa weil die Änderungen so tiefgreifend oder so zahlreich sind, daß ihre Tragweite nicht sofort übersehen werden kann, wäre zur Gewährung eines ausreichenden rechtlichen Gehörs auch nach einer mündlichen Verhandlung eine Mitteilung nach Regel 58 (4) EPÜ erforderlich.

7.2.1.2 Beispiele für die Sprachrichtung Deutsch-Französisch

7.2.1.2.1 Übersetzung von "rechtliches Gehör"

(9) rechtliches Gehör: principe du contradictoire (AN: NdeN)

Diese Prüfung muß mit großer Sorgfalt erfolgen, da der Grundsatz der Gewährung des **rechtlichen Gehörs** eines der wesentlichsten Prinzipien für ein gerechtes Verfahren darstellt.

Cet examen doit être effectué avec beaucoup de soin, car le respect du **principe du contradictoire** est essentiel pour assurer l'équité de la procédure.

(10) rechtliches Gehör: le droit des parties à être entendues[1] (NA: komplexe Nominalphrase)

Die mündliche Verhandlung ist daher die beste Form, das **rechtliche Gehör** den Parteien zu gewähren.

La procédure orale constitue par conséquent la meilleure façon de garantir le **droit des parties à être entendues**.

(11) rechtliches Gehör: le droit des parties à être entendues bzw. principe du contradictoire (NA: komplexe Nominalphrase bzw. NdeN)

Diese Zweifel hat die Kammer, weil nach ihrer Auffassung durch Regel 58 (4) EPÜ sichergestellt werden soll, daß im Einspruchsverfahren dem **Grundsatz des rechtlichen Gehörs**, der in Artikel 113 EPÜ niedergelegt ist, Rechnung getragen wird.

Si la Chambre se pose cette question, c'est parce qu'elle considère que la règle 58 (4) vise à garantir le respect durant la procédure d'opposition du **droit des parties à être entendues** (**principe du contradictoire**), posé à l'article 113 de la CBE.

1. Ein Terminologe aus dem Übersetzerdienst des Europäischen Patentamts bestätigte, daß es sich hier um einen Begriff handelt, für den es lange Zeit im Französischen keine kanonische Übersetzung gab, inzwischen gilt "le droit d'être entendues" als kanonische Übersetzung für "rechtliches Gehör".

(12) rechtliches Gehör: qu'elles seront suffisamment entendues (NdeN:Nebensatz)

> Findet eine mündliche Verhandlung nicht statt, so müssen andere Garantien geschaffen werden, die sicherstellen, daß den Parteien das **rechtliche Gehör** in ausreichendem Maße gewährt wird.
>
> S'il n'est pas organisé de procédure orale, il faut prévoir d'autres garanties pour que les parties puissent être assurées **qu'elles seront suffisamment entendues**.

(13) rechtliches Gehör: possibilités de se faire entendre (AN: komplexe Nominalphrase)

> In Zweifelsfällen wird auch die Kammer der Zustellung einer Mitteilung nach Regel 58 (4) immer den Vorzug geben, damit sichergestellt ist, daß das **rechtliche Gehör** den Parteien in jedem Falle ausreichend gewährt wird.
>
> En cas de doute, la chambre se prononcera toujours pour la signification de cette notification, afin de garantir dans tous les cas que les parties auront des **possibilités suffisantes de se faire entendre**.

7.2.1.2.2 Übersetzung von "Einspruchsbeschwerdeverfahren"

(14) Einspruchsbeschwerdeverfahren: procédure de recours engagée à l'encontre d'une décision rendue sur opposition (Komp: komplexe Nominalphrase)[1]

> Die Kammer ist jedoch der Ansicht, daß Regel 58 (4) nicht nur gemäß ihrem Wortlaut, sondern entsprechend ihrem Sinn und Zweck im **Einspruchsbeschwerdeverfahren** anzuwenden ist.
>
> La Chambre estime néanmoins que dans une **procédure de recours engagée à l'encontre d'une décision rendue sur opposition**, la règle 58 (4) ne doit pas uniquement s'appliquer à la lettre, mais qu'il peut également en être fait une application qui tienne compte de la signification qu'elle revêt et de la finalité qu'elle poursuit.

Besonderheiten der Übersetzungsäquivalenzen

Wie aus den Beispielen ersichtlich wurde, können die Übersetzungen eines Fachterminus durchaus variieren[2]. Um diese Tatsache zu berücksichtigen, wurden die Übersetzungsäquivalenzen für jede Form einzeln ermittelt und die jeweilige Frequenz bestimmt. Bei der Auswertung führt dies zu Problemen, wenn die kanonische Übersetzung immer verwendet wurde. Kommen beispielsweise in einem Text 10 Formen des Lemmas "Alkaligehalt" vor, die immer mit "teneur en alcali" übersetzt wurde (oder "Natriumgehalt", das mit "teneur en sodium" übersetzt wurde), so liegen in der Auswertung viele Fälle von Entsprechungen eines dt. Kompositums mit einer frz. Form des Typs *NPrepN* vor. Je nach Größe des ausgewerteten Textes kann dadurch der Ein-

1. Die Übersetzungen von "Einspruchsbeschwerdeverfahren" waren im ganzen Korpus einheitlich.
2. Dies ist jedoch nicht der Normalfall.

druck entstehen, daß die Entsprechung eines dt. Kompositums mit einer frz. Form des Typs *NPrepN* mit einer bestimmten Häufigkeit auftritt.

Deshalb wurden die Übersetzungsäquivalenzen unter zwei Aspekten ermittelt:

- Struktur des Übersetzungsäquivalents
- Übersetzung einzelner Nominalphrasen, d. h. es wird untersucht, ob eine gleichbleibende Übersetzung vorliegt, die wahrscheinlich als kanonische Übersetzung zu betrachten ist.

7.3 Strukturäquivalenzen

Für jede ermittelte Übersetzungsäquivalenz zwischen Formen A_{L1} und B_{L2} wird bestimmt, welche grammatische Struktur die Formen A_{L1} und B_{L2} haben. Ist beispielsweise *Beschwerdekammer* mit *chambre de recours* übersetzt worden, so ist ein dt. Kompositum im Frz. durch eine Form der Struktur *NdeN* wiedergegeben worden.

Aus drei Dokumenten wurden Strukturäquivalenzen erhoben, darunter ein Dokument, das viel fachspezifischen Wortschatz aus der Chemie enthält (Dok. 841 bzw. 843) und zwei Dokumente, deren Fachwortschatz etwa zu gleichen Teilen aus dem juristischen Bereich und aus der Chemie stammt (Dokumente 1087 bzw. 1089 und 2016 bzw. 2018).

Bei der Bestimmung der Strukturäquivalenzen wurden drei Fälle berücksichtigt:

- Die zielsprachliche Übersetzung der Form gehört einem Strukturtyp an, der für die Zielsprache extrahiert worden war, z. B.wenn ein dt. Nominalkompositum im Frz. einer Form der Struktur *NdeN* entspricht (wie in den Beispielen (1), (2), (3), (4), (6), (8), (9), (11) aus dem vorhergehenden Abschnitt).
- Die zielsprachliche Übersetzung ist eine Nominalphrase, deren Form aber nicht den extrahierten Mustern entspricht (wie in den Beispielen (10), (11), (13), (14) aus dem vorhergehenden Abschnitt).
- Die zielsprachliche Übersetzung ist strukturell ganz anders als die ausgangssprachliche Form, d. h. sie ist keine Nominalphrase (wie in den Beispielen (5), (7), (12) aus dem vorhergehenden Abschnitt).

7.3.1 Französisch-deutsche Strukturäquivalenzen

Es wurden 503 Formen der Struktur *NdeN* und 584 Formen der Struktur *NA* untersucht.

Die Ergebnisse werden gesondert für jedes Dokument angegeben. Es werden immer **nur die 3-4 häufigsten Arten** von Strukturäquivalenzen aufgeführt.

.

Tabelle 41. Frz.-dt. Strukturäquivalenzen aus Dok. 843

frz.: NdeN	Frequenz (%)	frz.: NA	Frequenz (%)
dt.: Kompositum	53,19	dt.: AN	56,32
dt.: einfaches Nomen	5,31	dt.: Kompositum	39,08
dt.: NgenN	25	dt.: Adverb	3,4
dt.: Bindestrichwort	9,04		
dt.: keine Nominalphrase	3,7	dt.: keine Nominalphrase	3,8

.

Tabelle 42. Frz.-dt. Strukturäquivalenzen aus Dok. 1089

frz.: NdeN	Frequenz (%)	frz.: NA	Frequenz (%)
dt.: Kompositum	51,33	dt.: AN	60,79
dt.: einfaches Nomen	5,34	dt.: Kompositum	14,77
dt.: NgenN	13,36	dt.: einfaches Nomen	13,06
dt.: NPN	9,62		
dt.: keine Nominalphrase	3,74	dt.: keine Nominalphrase	8,3

Tabelle 43. Frz.-dt. Strukturäquivalenzen aus Dok. 2018

frz.: NdeN	Frequenz (%)	frz.: NA	Frequenz (%)
dt.: Kompositum	63,28	dt.: AN	60,68
dt.: einfaches Nomen	7,03	dt.: Kompositum	29,91
dt.: NgenN	12,5	dt.: prädikatives Adjektiv	3,41
dt.: NPN	5,46		
dt.: keine Nominalphrase	4,6	dt.: keine Nominalphrase	4,27

7.3.2 Deutsch-französische Strukturäquivalenzen

Es wurden 811 Nominalkomposita und 582 Formen der Struktur *AN* untersucht.

Die Ergebnisse werden gesondert für jedes Dokument angegeben. Es werden immer **nur die 3-4 häufigsten Arten** von Strukturäquivalenzen aufgeführt.

Tabelle 44. Dt.-frz. Strukturäquivalenzen aus Dok. 841

dt.: Nominalkompositum	Frequenz (%)	dt.: AN	Frequenz (%)
frz.: NdeN	40,08	frz.: NA	65,78
frz.: N	19,43	frz.: N	8,77
frz.: NA	19,93	frz.: NdeNA	6,14
frz.: NaN	6,07		
frz.: keine Nominalphrase	6,88	frz.: keine Nominalphrase	7,85

Tabelle 45. Dt.-frz. Strukturäquivalenzen aus Dok. 1087

dt.: Nominalkompositum	Frequenz (%)	dt.: AN	Frequenz (%)
frz.: NdeN	34,98	frz.: NA	52,01
frz.: N	25,08	frz.:NdeN	4,03
frz.: NPrepN	9,42	frz.:NPrepN	3,58
frz.: NA	7,92		
frz.: keine Nominalphrase	16,49	frz.: keine Nominalphrase	24,01

Tabelle 46. Dt.-frz. Strukturäquivalenzen aus Dok. 2018

dt.: Nominalkompositum	Frequenz (%)	dt.: AN	Frequenz (%)
frz.: NdeN	36,63	frz.: NA	71,75
frz.: NA	21,97	frz.: NdeNA	5,21
frz.: N	20,37	frz.: NdeN	3,08
frz.: NdeNA	4,3		
frz.: keine Nominalphrase	12,82	frz.: keine Nominalphrase	13,91

7.3.3 Auswertung

Es wurde schon erwähnt, daß die Auswertung der Strukturäquivalenzen problematisch ist, weil jede Entsprechung zwischen zwei Fomen A_{L1}(Text1, Satz x) und B_{L2}(Text1, Satz y) einzeln, d. h. unabhängig von der Frequenz von A_{L1} und B_{L2} in den beiden Texten $Text1_{L1}$ und $Text1_{L2}$ gezählt wurden. Trotzdem sind aus den prozentualen Angaben in den Auswertungen der drei Textpaare einige Tendenzen ersichtlich.

Es fällt auf, daß weniger als 1 % der untersuchten frz. Nominalphrasen im Deutschen durch Nominalphrasen anderer, bisher nicht berücksichtigter Bildungsmuster (siehe Abschnitt 6.4.1.3 Die Extraktion von Strukturen, Seite 131) übersetzt sind, d. h. die für das Deutsche ausgewählten Bildungsmuster sind der Aufgabe angemessen.

Die dt. Nominalkomposita wurden im Frz. in 20 % der Fälle durch Nominalphrasen anderer, nicht berücksichtigter Bildungsmuster (siehe Abschnitt 5.4.3 Verwendete Suchmuster, Seite 87) übersetzt. Die dt. Nominalphrasen der Form AN wurden im Frz. in 9 % der Fälle durch Nominalphrasen anderer, nicht berücksichtigter Bildungsmuster übersetzt. Das bedeutet, daß selbst das beste Verfahren zur automatischen Bestimmung von Übersetzungsäquivalenzen in diesen Fällen kein korrektes Ergebnis hätte liefern können. Dasselbe gilt für die Fälle, in denen eine Nominalphrase in der einen Sprache in der anderen Sprache nicht durch eine Nominalphrase übersetzt wurde.

Bei der **Übersetzungsrichtung Französisch-Deutsch** ist festzustellen, daß im Deutschen:

- nur selten *nicht* durch eine Nominalphrase übersetzt wird (Werte zwischen 3,7 und 8,3 %),
- die Nominalphrasen in den meisten Fällen den Bildungsmustern KOMP, AN, NgenN, und NPN entsprechen[1],
- die frz. NdeN-Komposita am häufigsten mit Nominalkomposita (Werte zwischen 51,33 und 63 %) oder Nominalphrasen der Form NgenN (Werte zwischen 12,5 und 25 %) übersetzt werden,
- die frz. NA-Komposita am häufigsten mit Nominalphrasen der Form AN (Werte zwischen 56,32 und 60,67 %) oder Nominalkomposita (Werte zwischen 14,7 und 39,68 %) übersetzt werden,
- einfache Nomen keine bedeutende Rolle spielen (Werte zwischen 3 und 7 %).

Bei der **Übersetzungsrichtung Deutsch-Französisch** ist festzustellen, daß im Französischen:

- die Übersetzung der dt. Nominalkomposita am häufigsten durch eine Nominalphrase der Form NdeN erfolgt (Werte zwischen 35 und 40 %),
- die Häufigkeit der Übersetzung eines dt. Nominalkompositums durch ein einfaches Nomen nicht zu unterschätzen ist[2] (Werte zwischen 19,43 und 25,08 %),
- die Übersetzungen der dt. Nominalphrasen der Form AN am häufigsten durch NA-Komposita erfolgt (Werte zwischen 52,01 und 71,75 %), an zweiter Stelle stehen einfache Nomina, NdeN-Komposita und Nominalphrasen der Form NdeNA,
- die Anzahl der Fälle, in denen nicht durch eine Nominalphrase übersetzt wurde, sehr stark schwanken (Werte zwischen 6,88 und 24,01 %).

1. Diese Beobachtung deckt sich mit den formalen Beschreibungen der Fachtermini durch Drozd & Seibicke, Krause u. a., siehe Kapitel 6 Nominale Fachterminologie im Deutschen, Seite 107.

2. In den vorgestellten Arbeiten zur Form frz. Fachtermini (Abbildung 5.2, “Computerlinguistische Arbeiten”, auf Seite 65) wird diese Tatsache nur bei Sta 1995 erwähnt.

7.4 Übersetzungen einzelner Nominalphrasen

Die aus den Texten extrahierten Nominalphrasen wurden zur Ermittlung der Frequenz mit dem INTEX- bzw. dem CISLEX-System lemmatisiert. Für alle dt. und frz. Nominalphrasen mit einer Auftretenshäufigkeit größer 2 wurde untersucht, wie sie übersetzt wurden.

Die Ergebnisse werden exemplarisch an einem bilingualen Text (Dok. 841 bzw. 843) aufgeführt. Es werden ausgehend von einer Quellsprache die häufigsten Nominalphrasen und deren kanonische Übersetzungen mit den jeweiligen Auftretenshäufigkeiten (Frq.) angegeben.

Tabelle 47. Sprachrichtung frz.-dt. (Dok. 843/841)

frz.	Typ	Frq.	dt.	Typ	Frq.
état de la technique	NdeN	26	Stand der Technik	NgenN	26
bromure d'alkyle	NdeN	15	Alkylbromid	KOMP	14
activité inventive	NA	12	erfinderische Tätigkeit	AN	14
composé spiro	NN	10	Spiroverbindung	KOMP	11
homme du métier	NdeN	10	Fachmann	KOMP	10
essai comparatif	NA	10	Vergleichsversuch	KOMP	15
produit de départ	NdeN	8	Ausgangsstoff	KOMP	12
stabilisant à la lumière	NaN	6	Lichstabilisator	KOMP	8
bromure de méthyle	NdeN	6	Methylbromid	KOMP	6
objet de la demande	NdeN	5	Anmeldungsgegenstand	KOMP	5
squelette de base	NdeN	3	Grundgerüst	KOMP	5

Tabelle 48. Sprachrichtung dt.-frz. (Dok. 841/843)

dt.	Typ	Frq.	frz.	Typ	Frq.
Verbindung	N	30	composé	N	20
Anspruch	N	23	revendication	N	20
Stand der Technik	NgenN	26	état de la technique	NdeN	26
Vergleichsversuch	KOMP	15	essai comparatif	NA	10
Beschwerdeführerin	KOMP	14	réquerante	N	15
Alkylbromid	KOMP	14	bromure d'alkyle	NdeN	15
erfinderische Tätigkeit	AN	14	activité inventive	NA	12
Kammer	N	13	chambre	N	14
Entscheidung	N	13	décision	N	10
Ausgangsstoff	KOMP	12	produit de départ	NdeN	8
Spiroverbindung	KOMP	11	composé spiro	NN	10
Fachmann	KOMP	10	homme du métier	NdeN	10

Wenn die kanonischen Übersetzung nicht verwendet wird, so steht meist stattdessen ein Pronomen oder eine Art von Ellipse (z. B. "produit" statt "produit de départ" oder "Kammer statt "Beschwerdekammer").

7.5 Anwendungen

Es wurde kein vollautomatisches Verfahren für die Extraktion bilingualer Fachtermini angestrebt, da ein solches Verfahren meiner Ansicht nach zu viele Fehlerquellen in sich birgt. Die Möglichkeiten, **automatisch** aus der Menge der extrahierten Nominalphrasen diejenigen herauszufiltern, die echte Fachtermini sind, sind auf statistische Berechnungsverfahren begrenzt, deren Ergebnisse für eine Weiterverarbeitung meistens nicht zuverlässig genug sind. Das hat zur Folge, daß der Input eines automatischen Verfahrens zur Bestimmung von Entsprechungen Phänomene miteinander vergleicht, die einen unterschiedlichen Status haben: Fachtermini, potentielle Fachtermini und "gewöhnliche" Nominalphrasen. Darüber hinaus wird mit einem solchen Algorithmus postuliert, daß eine Übersetzungsäquivalenz zwischen den für beide Sprachen extrahierten Nominalphrasen existiert. Diese Annahme ist zwar für einen bestimmten Anteil der Nominalphrasen richtig, sie ist jedoch nicht in dem Maße generalisierbar, wie man erwarten könnte. In den vorhergehenden Kapiteln wurde unter verschiedenen Aspekten gezeigt, daß diese Annahme nur eingeschränkt gültig ist.

Der Nutzen der durchgeführten Untersuchungen liegt meines Erachtens mehr im Sinne einer Bereitstellung linguistischer Resourcen für den Übersetzer in Form eines Konkordanzprogramms.

Konkordanzprogramm

Das Programm erstellt aus den satzweise alignierten bilingualen Texten und der Liste der **häufigsten** potentiellen Fachtermini, die für jede Einzelsprache extrahiert wurden, Konkordanzen, die folgende Informationen bieten:

- Markierung der potentiellen Fachtermini in den bilingualen Texten
- Verweise auf weitere Textstellen (Konkordanzen für einzelne potentielle Fachtermini)
- Verweise auf potentielle Fachtermini mit gemeinsamen Grundwörtern

Die angebotenen Informationen werden hier aus Gründen der Übersichtlichkeit einzeln dargestellt.

7.5.1 Bilingualer Text mit Markierung der potentiellen Fachtermini

Die Ausgabe der Sätze und der potentiellen Fachtermini könnte in etwa die folgende Form[1] haben:

1. Im dt. Text sind Nominalkomposita durch Fettdruck, *AN*-Phrasen durch Kursivdruck und Phrasen anderer Bildungsmuster durch Unterstreichung gekennzeichet, im frz. Text sind *NdeN*-Phrasen durch Fettdruck, *AN*-Phrasen kursiv und Phrasen anderer Bildungsmuster durch Unterstreichung gekennzeichnet.

Tabelle 49. Beispiel eines parallelen Texts

dt. Text	frz. Text
II. Gegen diese Erteilung des *europäischen Patents* hat die Einsprechende am 11. August 1982 Einspruch eingelegt und den **Widerruf** des Patents wegen *mangelnder Neuheit* beantragt. Die Begründung wurde unter anderem auf *neue* ***Entgegenhaltungen*** gestützt.	II. Le 11 août 1982, la requérante a fait opposition à ce *brevet européen*, et en a demandé la révocation pour **défaut de nouveauté**, en faisant valoir notamment de *nouvelles antériorités*.
III. Durch Entscheidung vom 13. Oktober 1983 hat die **Einspruchsabteilung** den Einspruch zurückgewiesen. Die **Zurückweisung** wurde im wesentlichen damit begründet, daß der Gegenstand des **Patentanspruchs** 1 sich in zwei *wichtigen Merkmalen* von dem relevanten Stand der Technik, d.h. DE-A-2 442 240 (1) und US-A-4 016 245 (6) unterscheide und daher neu sei.	III. Par décision en date du 13 octobre 1983, la **Division d'opposition** a rejeté l'opposition, au motif essentiellement que **l'objet de la revendication** 1 présentait deux *caractéristiques importantes* qui n'étaient pas contenues dans l'état pertinent de la technique, à savoir en l'occurrence dans le *document allemand* A-24 42 240 (document (1)) et dans le *document américain* A-4 016 245 (document (6)), et devait par conséquent être considéré comme nouveau.
Es gebe auch im Stand der Technik keine Anhaltspunkte, die die Verwendung von **Hexamethylendiamin** bei der Herstellung von Zeolithen naheliegend erscheinen lassen.	Rien dans l'**état de la technique** ne permettait d'affirmer que **l'utilisation d'hexaméthylènediamine** dans la **préparation de zéolites** s'imposait à l'évidence.
Es sei als überraschend anzusehen, daß durch das Verfahren des **Streitpatents** direkt ein *alkalifreier Zeolith* hergestellt werden kann.	Le procédé revendiqué dans le brevet en litige permettait de préparer directement une zéolite sans alcali, possibilité qui devait être considérée comme inattendue.
Die anderen **Entgegenhaltungen** seien im **Prioritätsintervall** veröffentlicht und daher - da die Priorität zu Recht beansprucht sei - nicht zu berücksichtigen; denn **Prioritätsverlust** trete nicht dadurch ein, daß die **Analysenergebnisse** in den Beispielen der **Patentschrift** einerseits und den **Prioritätsunterlagen** andererseits nicht identisch seien.	Les autres antériorités avaient été publiées durant le **délai de priorité**, et ne pouvaient donc être prises en considération, puisque la priorité avait été revendiquée à juste titre; en effet, le fait que les **résultats d'analyses** ne soient pas identiques dans les exemples du **fascicule de brevet** d'une part et dans le texte du **document de priorité** d'autre part n'entraînait pas la perte du **droit de priorité**.

7.5.2 Konkordanzen für einzelne Fachtermini

Beispiele für solche Konkordanzen wurden gezeigt bei der Untersuchung der frz. Termini *décision de revocation, activité inventive* und *maintien du brevet* sowie der dt. Termini *rechtliches Gehör* und *Einspruchsbeschwerdeverfahren* (siehe Abschnitt 7.2.1 Ermittlung der Äquivalenzen, Seite 147).

Aus den lemmatisierten Texten können auch die Kontexte einzelner Termini automatisch bestimmt werden. So kann beispielsweise untersucht werden, mit welchen Präpositionen oder mit welchen Verben ein Terminus präferentiell verwendet wird.

Beispiele typischer Verwendungskontexte

Tabelle 50. Kontexte von “procédure orale”

Art des Kontexts	Beispiel
Präpositionen	dans une procédure orale
	avant la procédure orale
	lors d’une procédure orale
Verben	une procédure orale s’est tenue
	une procédure orale s’est déroulée
	la procédure orale a relevé
	comparaître à une procédure orale
	être représenté dans une procédure orale
	prendre part à une procédure orale
	interrompre une procédure orale
	demander de recourir à une procédure orale
	organiser une procédure orale
nominale Kontexte	la tenue d’une procédure orale
	l’interruption d’une procédure orale
	le procès-verbal d’une procédure orale
Kontexte mit präpositionaler Funktion	au cours d’une procédure orale
	après la tenue d’une procédure orale
	à l’issue d’une procédure orale
	au terme d’une procédure orale
	à la suite de la procédure orale

Tabelle 51. Kontexte von “mündlicher Verhandlung”

Art des Kontexts	Beispiel
Präpositionen	in einer mündlichen Verhandlung
	vor einer mündlichen Verhandlung
	während einer mündlichen Verhandlung
	nach einer mündlichen Verhandlung
Verben	eine mündliche Verhandlung findet statt
	die mündliche Verhandlung hat ergeben
	zu einer mündlichen Verhandlung erscheinen
	in einer mündlichen Verhandlung vertreten sein
	teilnehmen an einer mündlichen Verhandlung
	eine mündliche Verhandlung unterbrechen
	eine mündliche Verhandlung beantragen
nominale Kontexte	die Durchführung einer mündlichen Verhandlung
	die Beteiligten einer mündlichen Verhandlung
	die Unterbrechung einer mündlichen Verhandlung
	die Niederschrift über eine mündliche Verhandlung
Kontexte mit präpositionaler Funktion	bei Abschluß einer mündlichen Verhandlung
	am Schluß einer mündlichen Verhandlung
	am Ende einer mündlichen Verhandlung
	im Anschluß an eine mündliche Verhandlung

7.5.3 Zusammenstellung von Termini mit gemeinsamen Konstituenten

Aus den lemmatisierten[1] Formen der potentiellen Fachtermini können automatisch die Termini zusammengestellt werden, denen mindestens ein bedeutungstragendes Wort - in diesem Fall sind das Nomina, Adjektive und Partizipen - gemein ist. Ein solches Netzwerk von Termini bietet zusammen mit den Verweisen auf die entsprechenden Fundstellen im Korpus einen inhaltlichen und lexikalischen Überblick.

Diese Anordnung ist besonders auch für das Frz. interessant, weil damit das Problem der Abgrenzung der zweigliedrigen Komposita von den mehrgliedrigen zumindest teilweise umgangen werden kann. Es werden alle Vorkommen der extrahierten Bildungsmuster aufgelistet, auch wenn deren Status im einzelnen nicht mit Sicherheit geklärt werden kann.

Die folgende Graphik zeigt in Auszügen am Beispiel der Nomina *Patent* bzw. *brevet*, wie die Gruppierung zu einem solchen Netzwerk aussehen kann. Innerhalb des Netzwerks sind die Formen nach ihren Bildungsmustern angeordnet.

1. Die Lemmatisierung der dt. Formen schließt die Segmentierung durch das CISLEX-System ein.

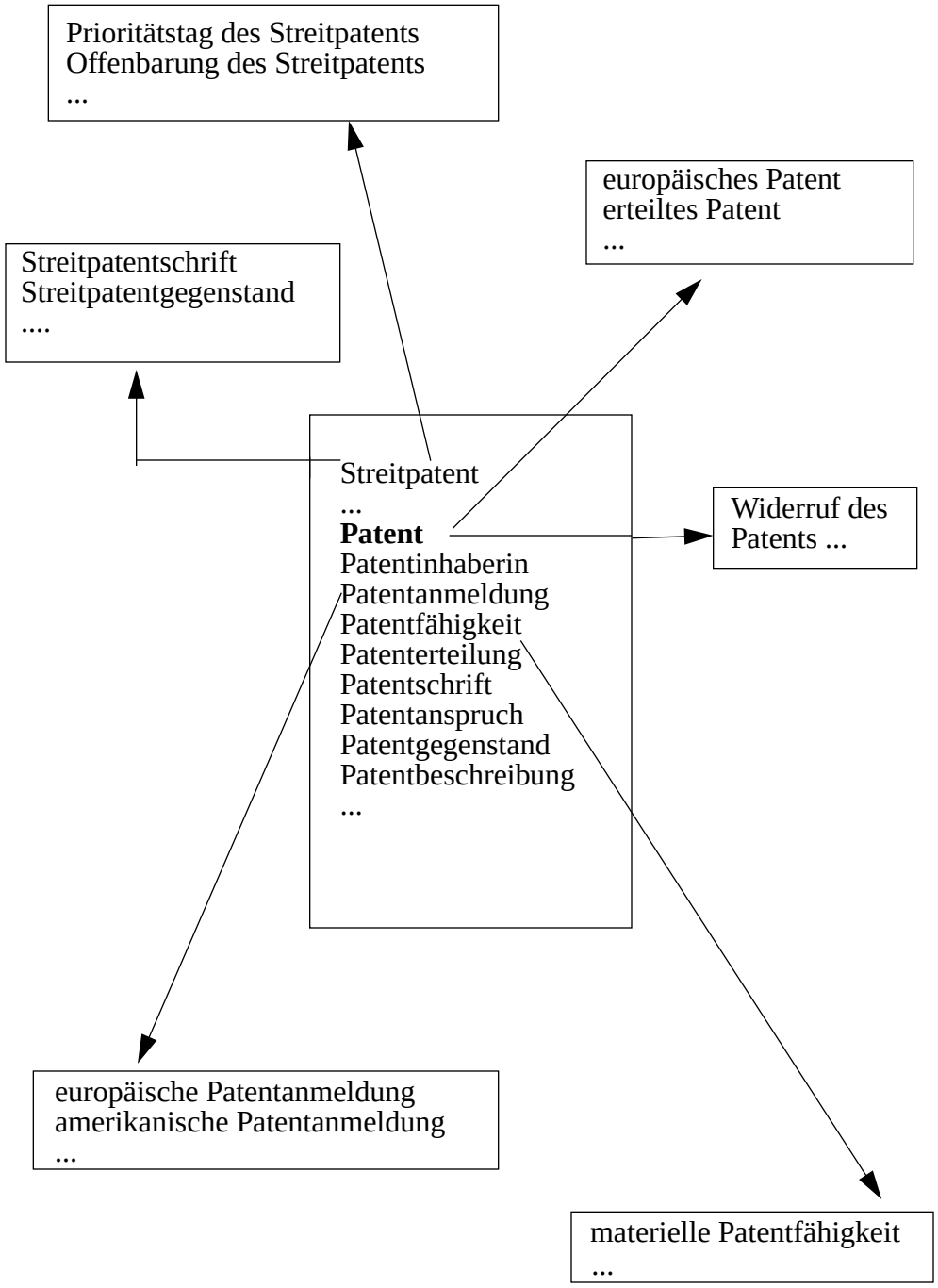


Abbildung 1: Ausschnitt aus einem Netzwerk von Termini für das Augangswort “Patent” (die Verweise auf die Fundstellen im Korpus sind hier nicht abgebildet)

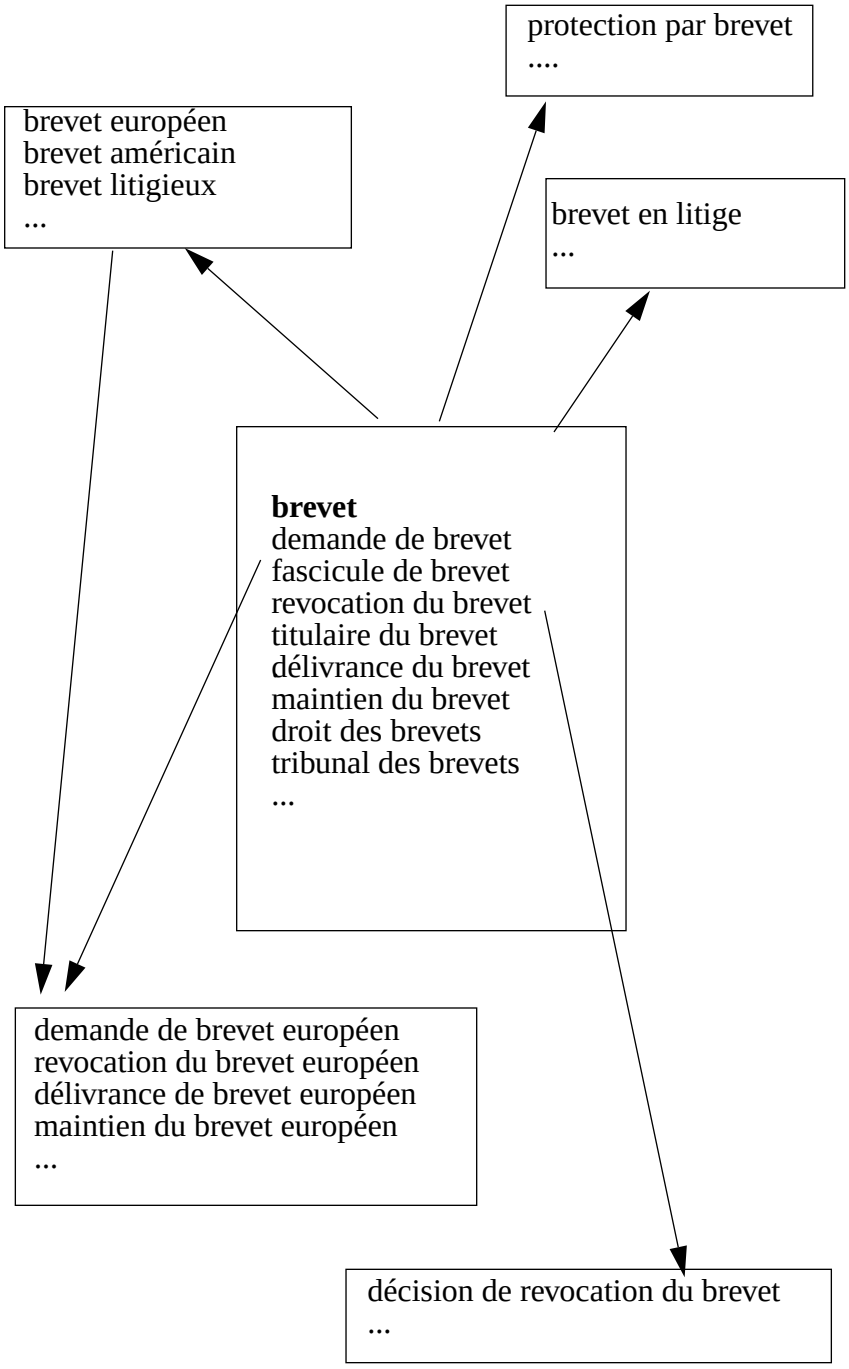


Abbildung 2: Ausschnitt aus einem Netzwerk von Termini für das Ausgangswort “brevet” (die Verweise auf die Fundstellen im Korpus sind hier nicht abgebildet)

8 Zusammenfassung und Ausblick

In dieser Arbeit wurde ein Verfahren für die Extraktion deutscher und französischer Termini aus einem Fachtextkorpus entwickelt. Die Ergebnisse werden in Form eines Konkordanzprogramms dargestellt, das als Hilfsmittel für Übersetzer verwendet werden kann.

Literatur aus der Fachsprachenforschung, der Übersetzungswissenschaft und der Computerlinguistik waren der Ausgangspunkt für die vorliegende Arbeit. Praxisorientierte Studien geben Hinweise darauf, welche Anforderungen an effiziente Werkzeuge zur Unterstützung des Übersetzungsprozesses gestellt werden. Die computerlinguistischen Verfahren wurden unter methodischen und linguistischen Aspekten untersucht. In der Vorgehensweise, d. h. in der Aufteilung der Aufgabe in einzelne Arbeitschritte, sind sich die Verfahren sehr ähnlich: fast alle verwenden Satzalignierung und Pattern-Matching-Techniken für die Extraktion der potentiellen Fachtermini. Die Ermittlung der Übersetzungsäquivalenzen basiert auf der Hypothese einer strukturellen Entsprechung. Man nimmt an, daß den Nominalphrasenmustern, die für die Quellsprache extrahiert werden, in der Zielsprache eine begrenzte Zahl von Nominalphrasenmustern entsprechen, teilweise wird sogar von einer 1:1-Entsprechung von Fachtermini in mehreren Sprachen ausgegangen. Die Gültigkeit dieser Hypothesen sollte am Testkorpus überprüft werden, um zu entscheiden, ob ein vollautomatisches Verfahren zur Ermittlung der Übersetzungsäquivalenzen sinnvoll ist. Zunächst bestand die Aufgabe darin, Verfahren zur Satzalignierung bereitzustellen und die einzelsprachlichen Bildungsmuster für Nominalphrasen zu definieren und aus den Texten zu extrahieren.

Zwei Satzalignierungsalgorithmen wurden für alle Sprachrichtungen des dt.-engl.-frz. Ausgangskorpus getestet. Der Church-Gale-Algorithmus zeigte die besseren Ergebnisse, unter der Voraussetzung, daß die einzelnen Abschnitte in den Texten nicht zu lang sind und im Vorfeld eine optimale Satzendeerkennung erfolgt ist. Hier erwies sich die gute und durchgängige Strukturierung des Korpus als großer Vorteil. Es ist zu prüfen, ob bei einer Anwendung der Alignierungsalgorithmen auf andere Textsorten vergleichbare Ergebnisse erzielt werden.

Französische potentielle Fachtermini wurden als Nominalphrasen bestimmter Bildungsmuster definiert und mit Hilfe des INTEX-Systems aus den Texten extrahiert. Das INTEX-System erwies sich als leistungsstarkes Werkzeug für diesen Zweck. Es ist damit zu rechnen, daß sich die linguistische Korrektheit der extrahierten Sequenzen noch erhöht, wenn alle in INTEX vorgesehenen Möglichkeiten zur Lösung von Ambiguitäten eingesetzt werden. Das größte Problem bestand darin, die Bildungsmuster für Nominalphrasen festzulegen, die bei der Extraktion der potentiellen Termini zu berücksichtigen sind. Die Frage, ob komplexe Nominalphrasen in Teilphrasen zu zerlegen sind und wie diese Zerlegung auszusehen hat, kann nicht allgemeingültig im voraus geklärt werden, sondern erfordert eine intellektuelle Nachbearbeitung für jeden Einzelfall. Hier erwies sich die Zusammenstellung von potentiellen Fachtermini mit gemeinsamen Konstituenten zusammen mit Verweisen auf die jeweiligen Fundstellen

als sinnvoll. In einigen Arbeiten aus der Literatur wurde dieses Zerlegungsproblem durch Frequenzinformation gelöst, d. h. wenn die Teile einer komplexen Nominalphrase häufiger auftreten als die komplexe Nominalphrase, wird davon ausgegangen, daß die Zerlegung korrekt ist. Dieser Ansatz könnte an einem größeren Auschnitt des EBK-Korpus überprüft werden.

Die Literatur zur Bildung der deutschen Termini bezieht sich hauptsächlich auf die einfachen und komplexen Nomina (Nominalkomposita) und Adjektiv-Nomen-Verbindungen. Nomina wurden durch das CISLEX-System mit nahezu 100% Korrektheit in den Texten erkannt und im Falle von Komposita richtig segmentiert. Die aufgetretenen Ambiguitäten sind auf eine sehr differenzierte morphologische Kodierung des CISLEX zurückzuführen und konnten in fast allen Fällen vernachlässigt werden. Die Extraktion der Adjektiv-Nomen-Verbindungen führte zu zufriedenstellenden Ergebnisse. Andere in der Literatur aufgeführte Arten von Mehrwortphrasen wurden mit einem Perl-Programm extrahiert. Dieser Ansatz ist jedoch nicht ausreichend angesichts der großen Zahl äußerst langer und komplexer Nominalphrasen, die für Patenttexte typisch ist. Eine angemessenere Lösung erfordert zumindest eine partielle syntaktische Analyse. Als Fachtermini spielten diese Mehrwortphrasen eine geringe Rolle. Es ist zu prüfen, ob diese Beobachtung bei der Untersuchung eines größeren Korpus bestätigt werden kann.

Ein Ziel der kontrastiven Auswertung bestand darin, die Hypothesen zu überprüfen, auf denen einige Ansätze zur automatischen bilingualen Terminologieextraktion basieren. Die These der 1:1-Entsprechung von Fachtermini konnte durch eine Vielzahl von Textbeispielen widerlegt werden. Außerdem stellte sich heraus, daß die Hypothese der strukturellen Entsprechungen vor allem für die Sprachrichtung Französisch-Deutsch nur eingeschränkt gilt. In manchen Texten waren bis zu 20 % der deutschen potentiellen Termini im Französischen entweder durch Nominalphrasen komplexerer Bildungsmuster oder nicht durch Nominalphrasen übersetzt. Diese Beobachtung läßt sich zumindest zum Teil mit den unterschiedlichen Wortbildungsverfahren der beiden Sprachen erklären. Ein deutsches Nominalkompositum ist problemlos automatisch erkennbar, während im Französischen die einzelnen Bestandteile eines Kompositums normalerweise orthographisch voneinander getrennt sind und erst durch eine morphosyntaktische und lexikalische Analyse zusammengeführt werden können. Dieses Ergebnis wirft die Frage auf, inwieweit die Studien, in denen Terminologie nicht nur automatisch extrahiert sondern auch aligniert wird, von den betrachteten Sprachpaaren abhängig sind. Eine bilinguale Alignierung vollständiger Nominalphrasen könnte zwar u. U. erfolgreicher sein, wäre aber mit dem Problem verbunden, daß zum einen nicht alle Nominalphrasen Fachtermini bezeichnen und zum anderen in den Texten zu viele verschiedene Nominalphrasen vorkommen, deren jeweilige Textfrequenzen niedrig ausfallen würden. Aus diesen Gründen wurde kein vollautomatisches Verfahren zur Ermittlung von Übersetzungsäquivalenzen angestrebt, sondern eine Darstellung der Ergebnisse in Form eines Konkordanzprogramms vorgezogen. Dieses Programm erstellt aus den satzweise alignierten bilingualen Texten und der Liste der extrahierten potentiellen Fachtermini verschiedene Arten von Konkordanzen. Diese ermöglichen es, Übersetzungsbeispiele für einzelne Fachtermini aufzuzeigen und typische Verwen-

dungskontexte zu bestimmen. Die Gesamtdarstellung der bilingualen Texte mit einer Markierung der potentiellen Fachtermini liefert Hinweise darauf, welche lexikalischen Einheiten terminologischen Charakter haben könnten. Aus potentiellen Fachtermini mit gemeinsamen Konstituenten und Verweise auf die entsprechenden Fundstellen im Korpus wird eine netzwerkartige Struktur erzeugt, die einen inhaltlichen und lexikalischen Überblick über Teilbereiche eines Fachgebiets geben können. In einer konkreten Anwendung könnte geklärt werden, ob darüber hinaus noch weitere linguistische Informationen aus den Texten für einen Übersetzer bereitgestellt werden sollten.

Die Untersuchung hatte zum Ziel, ein allgemeines Verfahren für die mehrsprachige Terminologieextraktion zu entwickeln. Als Testkorpus stand ein Korpus aus der Patentdokumentation zur Verfügung. Es wäre interessant, anhand eines Korpus aus einem anderen Bereich zu klären, ob die durchgeführte linguistische Beschreibung der Fachtermini Spezifika der Patentdokumentation enthält und ob die Fachtermini in anderen Fachgebieten mehr Variationen aufweisen.

9 Bibliographie

Arntz, Reiner; **Picht**, Heribert: Einführung in die übersetzungsbezogene Terminologiearbeit, Hildesheim: Olms, 1982.

Benveniste, Emile: Formes nouvelles de la composition nominale. *In*: Problèmes de la linguistique générale 2, Paris: Gallimard, 1974.

Bergenholtz, Henning; **Tarp**, Sven: Mehrworttermini und Kollokationen in Fachwörterbüchern. *In*: Schaeder, Burkhard; Bergenholtz, Henning (Hrsg.): Fachlexikographie, Tübingen: Narr, 1994.

Blank, Ingeborg: Etude des constructions syntaxiques en vue d'un traitement automatique. *In*: Les Cahiers du CRISS (Centre de Recherche d'Informatique appliquée aux Sciences Sociales), Grenoble, 1987, p. 1-52.

Blank, Ingeborg: Sentence alignment: methods and implementations. *In*: Traitement automatique des langues, Numéro special: Traitements probabilistes et corpus, Vol. 36, 1995, S. 81-99.

Blank, Ingeborg: Utilisation d'INTEX dans un projet d'extraction de terminologie. *In*: Actes des Premières Journées INTEX, LADL, Paris, 1996, S. 103-110.

Blatt, Achim: The EURAMIS Project. Working Paper, European Commission's Translation Service, Luxembourg, 1995.

Bourigault, Didier: Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. *In*: Actes de COLING, Nantes, 1992, S. 977-981.

Bourigault, Didier: An Endogeneous Corpus-Based Method for Structural Noun Phrase Disambiguation. *In*: Proc. of the Sixth Conference of the European Chapter of the Association for Computational Linguistics, Utrecht, 1993, S. 81-86.

Bourigault, Didier: LEXTER, un Logiciel d'EXtraction de TERminologie: Application à l'acquisition des connaissances à partir de textes. Diss. Ecole des Hautes Etudes en Sciences Sociales, Paris, 1994.

Bourigault, Didier: LEXTER, a Natural Language Processing Tool for Terminology Extraction. *In*: Proc. of EURALEX, Göteborg, 1996.

Brown, Peter; **de Souza**, Peter V.; **Mercer**, Robert L.; **Della Pietra**, Vincent J.; **Lai**, Jenifer C.: Class-based n-gram Models of Natural Language. *In*: Computational Linguistics, Volume 18, Number 4, 1992, S. 467-479.

Brown, Peter; **Cocke**, J.; **Della Pietra**, S.; **Della Pietra**, V.; **Jelinek**, F.; **Lafferty**, John D.; **Mercer**, R.; **Roossin**, P.: A Statistical Approach to Language Translation. *In*: Computational Linguistics, Vol. 16, Number 2, 1990, S. 79-85.

Calzolari, Nicoletta; **Bindi**, Remo: Acquisition of Lexical Information from a Large Textual Italian Corpus. *In*: Proc. of COLING, Helsinki, 1990, S. 54-59.

Catizone, Roberta; **Russell**, Graham; **Warwick**, Susan: Deriving Translation Data from Bilingual Texts. *In*: Zernik, U. (Hrsg.): Proc. of the First Lexical Acquisition Workshop, Detroit, 1989.

Church, Kenneth W.; **Hanks**, P.: Word Association Norms, Mutual Information and Lexicography. *In*: Computational Linguistics, Volume 16, Number 1, 1990, S. 22-26.

Church, Kenneth W. ; **Gale**, William A.: Concordances for Parallel Text. *In*: Proc. Seventh Annual Conference of the UW Centre for the New OED and Text Research Using Corpora, Oxford, 1991, S. 40-62.

Church, Kenneth W.; **Mercer**, Robert L.: Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *In*: Computational Linguistics, Volume 19, Number 1, 1993, S. 1-24.

Church, Kenneth W.; **Dagan**, Ido: Termight: Identifying and translating Technical Terminology. *In*: Proc. of the 4th Conference on Applied Natural Language Processing, Stuttgart, 1994, S. 34-40.

Coseriu, Eugenio: Structure lexicale et enseignement du vocabulaire. *In*: Actes du premier colloque international de linguistique appliquée, Nancy, 1966, S. 175-222.

Dagan, Ido; **Itai**, Alon: Word sense disambiguation using a second language monolingual corpus. *In*: Proc. of the 31th Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993, S. 1-39.

Dagan, Ido; **Church**, Kenneth W.; **Gale**, William: Robust bilingual word alignment for machine aided translation. *In*: Proc. of the Workshop on Very Large Corpora, 1994, S. 1-8.

Daille, Béatrice: Approche mixte pour l´extraction de terminologie: statistique lexicale et filtres linguistiques. Thèse de doctorat en informatique fondamentale, Université Paris VII, 1994.

Daille, Béatrice: Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *In*: Traitement automatique des langues, Vol. 36, 1995, S. 101-118.

Debili, Fathi; **Sammouda**, Elyès: Appariement des phrases de textes bilingues Francais-anglais et francais-arabe. *In*: Actes de COLING, Nantes, 1992, S. 518-524.

Dederding, H.-M.: Wortbildung, Syntax, Text. Nominalkomposita und entsprechende syntaktische Strukturen in deutschen Patent- und Auslegeschriften, Erlangen: Palm & Enke, 1982.

Dempster, A. P.; **Laird**, N. M.; **Rubin**, D. B.: Maximum Likelihood from incomplete data via the EM algorithm. *In*: Journal of the Royal Statistic Society, 39 (B), 1977, S. 1-38.

Downing, Pamela: On the creation and use of English compound nouns. *In*: Language, 53:4, 1977, S. 810-842.

Drozd, Lubomir; **Seibicke**, Wilfried: Deutsche Fach- und Wissenschaftssprache, Wiesbaden: Brandstetter, 1973.

Dunning, Ted: Accurate Methods for the Statistics of Surprise and Coincidence. *In*: Computational Linguistics, Vol. 19, Number 1, 1993, S. 61-74.

Eijk, Pim van der: Automating the acquisition of Bilingual Terminology. *In*: Proc. of the Meeting of the European Chapter of the Association for Computational Linguistics, Utrecht, 1993, S. 113-119.

Felber, Helmut; **Budin**, Gerhard: Terminologie in Theorie und Praxis, Tübingen: Narr, 1989.

Fluck, Hans-Rüdiger: Fachdeutsch in Wissenschaft und Technik, Heidelberg: Groos, 1984.

Fluck, Hans-Rüdiger: Fachsprachen, Tübingen: Francke, 3. Auflage, 1985.

Gale, William; **Church**, Kenneth W.; **Yarowsky**, David: Estimating Upper and Lower Bounds on the Performance of Word-Sense Disambiguation Programs. *In*: Proc. of the 30th Annual Meeting of the Association for Computational Linguistics, Delaware, 1992, S. 249-256.

Gale, William A.; **Church**, Kenneth W.: A program for aligning sentences in bilingual corpora. *In*: Proc. of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, 1991.

Gale, William A.; **Church**, Kenneth W.: A program for aligning sentences in bilingual corpora. *In:* Computational Linguistics, 19, Number 1, 1993, S. 75-102.

Gaussier, Eric: Extraction automatique de lexiques bilingues par des méthodes statistiques. Thèse de doctorat en informatique fondamentale, Université Paris VII, 1995.

Gross, Gaston: Degrés de figement des noms composés. *In*: Langages, Vol. 90, 1988, S. 57-72.

Gross, Maurice: Sur les phrases figées complexes du français. *In*: Langue française, Nr. 77: Syntaxe des connecteurs, 1988.

Habert, Benoît; **Jacquemin**, Christian: Noms composés, termes, dénominations complexes: problématiques linguistiques et traitement automatiques. *In*: Traitement automatique des langues, Vol. 34, 1993, S. 5-41.

Henzler, Rolf G.: Information und Dokumentation, Berlin: Springer, 1992.

Hoffmann, Lothar: Kommunikationsmittel Fachsprache, Tübingen: Narr, 1985.

Hoffmann, Lothar (Hrsg.): Fachsprachen: Instrument und Optik, Leipzig: Verlag Enzyklopädie, 1987.

Isabelle, Pierre: Bi-textual aids for translators. *In*: Proc. of the Annual Conference of the UW Center for the New OED and Text Research, 1992.

Isabelle, Pierre: Translation Analysis and Translation Automation. *In*: Proc. of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation, Kyoto, 1993.

Jacquemin, Christian: Transformation des noms composés. Thèse de doctorat en Informatique Fondamentale, Université Paris VII, 1991.

Kay, Martin; **Röscheisen**, Martin: Text-Translation Alignment. Technical Report, Xerox Palo Alto Research Center, 1988.

Kay, Martin; **Röscheisen**, Martin: Text-Translation Alignment. In: Computational Linguistics, Volume 19, Number 1, 1993, S. 121-142.

King, Maghi: European Commission Translation Service: A case study. Working Paper, European Commission's Translation Service, Luxembourg, 1995.

Klavans, Judith; **Tzoukermann**, Evelyne: The BICORD system: Combining lexical information from bilingual corpora and machine readable dictionaries. *In*: Proc. of COLING, Helsinki, 1990, S. 174-179.

Krause, Jürgen (Hrsg.): Inhaltserschließung von Massentexten, Hildesheim: Olms, 1988.

Krause, Jürgen; **Womser-Hacker**, Christa (Hrsg.): Das Deutsche Patentinformationssystem, Köln: Heymanns, 1990.

Kupiec, Julian: An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora. *In*: Proc. of the 31th Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993, S. 17-22.

Landauer, Thomas K.; **Littman**, Michael L.: Fully Automated Cross-Language Document Retrieval Using Latent Semantic Indexing. *In*: Proc. of the Annual Confe-

rence of the UW Centre for the New OED and Electronic Text Research, Waterloo, Ontario, 1990.

Langer, Stefan: Selektionsklassen und Hyponymie im Lexikon. Centrum für Informations- und Sprachverarbeitung (CIS-Bericht 96-94), München, 1996.

Maier-Meyer, Petra: Lexikon und automatische Lemmatisierung. Centrum für Informations- und Sprachverarbeitung (CIS-Bericht 95-84), München, 1995.

Mathieu-Colas, Michel: Typologie des mots composés. Rapport technique numéro 7, Programme de Recherches Coordonnés "Informatique et Linguistique", Université Paris XIII, 1988.

Maxwell, Kerry: Automatic translation of English compounds: problems and prospects. Working Papers in Language Processing, Number 39, Univ. of Essex, 1992.

Möhn, Dieter; **Pelka**, Roland: Fachsprachen: Eine Einführung, Tübingen: Narr, 1984.

Mossmann, Yvan: Terminologistik: Eine neue Dimension der Terminologiearbeit. *In*: Fischer, Ingeborg; Freigang, Karl-Heinz; Mayer, Felix; Reinke, Uwe (Hrsg.): Sprachdatenverarbeitung für Übersetzer und Dolmetscher, Akten des Symposiums zum Abschluß des Saarbrücker Modellversuchs, Hildesheim: Olms, 1994, S. 21-29.

Motsch, W.: Zur Stellung der Wortbildung in einem formalen Sprachmodell. *In*: Studia Grammatica I, 31-50, 3. Aufl. 1966.

Mosteller, Frederick; **Wallace**, David: Inference and Disputed Autorship, Reading, Massachusetts: Addison Wesley, 1964.

Reinart, Sylvia: Terminologie und Einzelsprache, Frankfurt: Lang, 1993.

Reinhardt, Werner; **Köhler**, Claus; **Neubert**, Gunter : Deutsche Fachsprache der Technik, Hildesheim: Olms, 1992.

Revuz, Dominique: Dictionnaires et lexiques, méthodes et algorithmes. Thèse de doctorat, CERIL, Université Paris VII, 1991.

Salton, Gerard: Automatic Text processing, Reading, Massachusetts: Addison Wesley, 1989.

Sato, Satoshi; **Nagao**, Makoto: Toward memory-based Translation. *In*: Proc. of COLING, Helsinki, 1990, S. 247-252.

Schaeder, Burkhard; **Bergenholtz**, Henning (Hrsg.): Fachlexikographie, Tübingen: Narr, 1994.

Schamlu, F.: Patentschriften - Patentwesen: Eine argumentationstheoretische Analyse, München: Fink, 1985.

Schicht, Gabi: Probleme der Satzendeerkennung. Centrum für Informations- und Sprachverarbeitung (CIS-Bericht 94-81), München, 1994.

Schmitt, Peter A.: Der Translationsbedarf in Deutschland: Ergebnisse einer Umfrage. *In*: Mitteilungsbedarf für Übersetzer und Dolmetscher 39/5, 1993, S. 3-10.

Schmitt, Peter A.: Translationsorientierte Terminographie am PC. *In*: Fischer, Ingeborg; Freigang, Karl-Heinz; Mayer, Felix; Reinke, Uwe (Hrsg.): Sprachdatenverarbeitung für Übersetzer und Dolmetscher, Akten des Symposiums zum Abschluß des Saarbrücker Modellversuchs 1992, Hildesheim: Olms, 1994, S. 31-61.

Schneider, Christine: Automatische Indexierung und Syntaxanalyse, Hamburg: Buske, 1985.

Schwanke, Martina.: Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis, Berlin: Springer, 1991.

Seelbach, Dieter: Computerlinguistik und Dokumentation, München: UTB Verlag, 1975.

Silberztein, Max : Dictionnaires éléctroniques et reconnaissance lexicale automatique, Paris: Masson, 1993.

Simard, M.; **Foster**, G.; **Isabelle**, P.: Using Cognates to Align Sentences in Bilingual Corpora. *In*: Proc. of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 92), Montreal, Canada, 1992.

Sta, Jean-David: Comportement statistique des termes et acquisition terminologique à partir de corpus. *In*: Traitement automatique des langues, Vol. 36, 1995, S. 119-132.

Wallner, Margot: Untersuchung von Text-Alignment im Rahmen computergestützter Textübersetzung auf der Basis des Kay-Röscheisen Algorithmus. Diplomarbeit Fachhochschule München, Fachbereich Informatik, 1994.

Warwick, S.; **Hajic**, J.; **Russell**, G.: Searching on tagged corpora: linguistically motivated concordance analysis. Working Paper, ISSCO, Geneva, 1993.

Wittmann, Alfred: Patentdokumentation. *In*: Buder, Marianne; Rehfeld, Werner; Seeger, Thomas (Hrsg.): Grundlagen der praktischen Information und Dokumentation, München: Saur, 3. Aufl. 1991, S. 522-533.

Womser-Hacker, Christa: Gegenüberstellung von intellektueller und maschineller Recherche auf der Basis von Aufgaben des PADOK-Retrievaltests. PADOK-Arbeitsbericht, Universität Regensburg, 1986.

Wüster, Eugen: Internationale Sprachnormung in der Technik, besonders in der Elektrotechnik. Diss. 1931, 3. erg. Auflage, Bonn: Bouvier, 1970.

Wüster, Eugen: Die allgemeine Terminologielehre - ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. *In*: Proc. of the Third Congress of the “Association Internationale de Linguistique Appliquée” Copenhagen, Vol. III: Applied Linguistics - Problems and Solutions, Heidelberg: Groos, 1974, S. 640-655.

Zimmermann, H.: CTX- Ein Verfahren zur Texterschließung. BMFT-Forschungsbericht D 83-006, 1983.