Aus dem Institut für Medizinische Biometrie und Medizinische Informatik der
Albert-Ludwigs-Universität Freiburg i.Br.

Das MorphoSaurus-System
-
Lösungen für die linguistischen Herausforderungen des Information Retrievals in der Medizin

INAUGURAL - DISSERTATION
zur
Erlangung des Medizinischen Doktorgrades
der Medizinischen Fakultät
der Albert-Ludwigs-Universität
Freiburg i.Br.

vorgelegt **2007**
von **Philipp Daumke**
geboren in **Freiburg i.Br.**

Dekan:	Prof. Dr. med. Christoph Peters
1. Gutachter:	Prof. Dr. med. Stefan Schulz
2. Gutachter:	PD Dr. med. Elmar Kotter
Jahr der Promotion:	2008

Zusammenfassung

In der Medizin werden in großem Umfang textuelle Informationen sowohl in Wissenschaft als auch in Klinik und Praxis erzeugt und benötigt. Die Variationen der natürlichen Sprache stellen dabei eine enorme Herausforderung für rechnergestützte Anwendungen dar, die sich mit der Verarbeitung und dem Wiederfinden dieser Informationen beschäftigen.

In dieser Arbeit wird das MorphoSaurus-System als eine Anwendung zur Verarbeitung und zum Wiederfinden von medizinischen Dokumenten aus Klinik und Wissenschaft vorgestellt. Dieses System realisiert eine ein- und mehrsprachige Normalisierung von Dokumenten und zerlegt dabei Wörter in ihre semantischen Grundbestandteile. Derzeit werden sechs europäische Sprachen von dem System unterstützt. Durch die sprachliche Normalisierung stellt MorphoSaurus Lösungen für zahlreiche sprachliche Variationen auf morphologischer, syntaktischer und lexikosemantischer Ebene bereit. Es wurde ursprünglich für die mehrsprachige Dokumentenrecherche (engl. *Cross Language Information Retrieval*) entwickelt, lässt sich jedoch auch in zahlreichen weiteren Anwendungen zur natürlichen Sprachverarbeitung einsetzen. In dieser Arbeit werden Experimente zur einsprachigen Dokumentenrecherche, zur automatischen Textkategorisierung und zur Übersetzung medizinischer Terme dargestellt sowie die Integration des Recherchesytems in das Informtionssystem der Hautklinik beschrieben.

In den Experimenten zur Dokumentenrecherche zeigen sich insbesondere Verfahren erfolgreich, in denen das MorphoSaurus-System mit klassischen regelbasierten Verfahren zur Stammformbildung kombiniert wird. Die Experimente zur automatischen Textkategorisierung verdeutlichen die Überlegenheit des statistischen Verfahrens gegenüber dem heuristischen Verfahren. In den Experimenten zur automatischen Übersetzung biomedizinischer Terme wurden in der deutsch-englischen Übersetzung über 3/4 aller Terme korrekt oder sinnverwandt übersetzt. Die durchgeführte Fehleranalyse verdeutlicht, dass die Ursachen für fehlerhafte Übersetzungen zum überwiegenden Teil in der fehlenden Abdeckung der MorphoSaurus-Lexika und nur selten in einer Schwäche der vorgestellten Methode liegen. Die Einbindung des MorphoSaurus-Systems in das Informationssystem der Hautklinik wurde von den Ärzten sehr positiv aufgenommen. Das System soll in den kommenden Monaten um zusätzliche Funktionalitäten wie Einbindung von Labordiagnosen erweitert werden.

Curriculum Vitae

Ausbildung	2002–derzeit	**Abt. Medizinische Informatik**, Universitätsklinikum Freiburg, Deutschland Dissertation Forschungsbereiche: ➢ Automatische Textkategorisierung biomedizinischer Dokumente ➢ Biomedizinisches Textretrieval ➢ Text Mining
	2000–derzeit	**Fern-Universität Hagen**, Hagen, Deutschland Studium der Informatik Angestrebter Abschluss: Bachelor of Science, Abschluss: 2008
	1997–2004	**Universität Freiburg/** Krankenhaus Konstanz, Deutschland **University of Queensland**/ Mater Hospital, Brisbane, Australien **University of Melbourne**/ Austin Hospital, Melbourne, Australien **University of Tasmania**/ Royal Hobart Hospital, Hobart, Australien **Albert-Ludwigs-Universität Freiburg**, Studium der Medizin Staatsexamen: gut (1,66) (top 10%) Physikum: sehr gut (1,0) (top 1%)
	1987–1996	**Kolleg St. Sebastian**, Stegen. Abitur, Note: 1,0
Berufserfahrung	04/07-derzeit	**Geschäftsführender Gesellschafter** der Averbis GmbH
	09/05–12/05	**Siemens Medical Solutions**, Erlangen, Deutschland Praktikum in der Abteilung ‚Software Components and Workstations Marketing'
	10/02 09/02 09/01–10/01 02/00–03/00	**Universitätsklinikum, Freiburg**, Deutschland Famulatur im Fach ‚Anästhesie' **LIDG-Hospital**, Oban, Schottland Famulatur im Fach ‚Chirurgie' **Somerset Hospital**, Kapstadt, Südafrika Famulatur im Fach ‚Innere Medizin' **Prof. Dr. Habermeyer**, Heidelberg, Deutschland Famulatur im Fach ‚Schulterchirurgie'
	10/96–09/97	**St. Josefshospital**, Freiburg, Zivildienst
Publikationen	2003-derzeit	Über 20 Publikationen in internationalen Zeitschriften in Medizinischer Informatik und Information Retrieval sowie zahlreiche internationale Fachvorträge
Auszeichnungen/ Stipendien	2007 2005–2006 2000–derzeit 2003	Johann-Peter-Süßmilch-Medaille der GMDS e.V. Karl-Steinbuch-Stipendium Stipendiat bei *e-fellows.net* DeAN-Qantas-Reisestipendium
Weiteres		Sprachen: Englisch, Französisch Snowboard-Ausbilder des Deutschen Skiverbandes Sportarten: Snowboard, Mountain-Biking Reisen: Australien, Schottland, Südafrika, USA

Wissenschaftliche Veröffentlichungen

Konferenzen und Zeitschriften

2007

P. Daumke, K. Markó, J. Paetzold, M. Müller: *Biomedical Data Mining in a Hospital Information System*. Accepted for MedNet 2007 in Leipzig, Germany.

K. Marko, **P. Daumke**, S. Schulz, R. Klar, U. Hahn: *Large-Scale Evaluation of a Medical Cross-Language Information Retrieval System*. In: K.A. Kuhn, J.R. Warren, T.-Y. Leong (Hrsg): MEDINFO 2007 - Proceedings of the 12th World Congress on Health (Medical) Informatics - Building sustainable Health Systems, Studies in Health Rechnology and Informatics 129, pp. 529-534. Brisbane, Australia, 2007, Amsterdam: IOS Press.

M. Müller, K. Markó, **P. Daumke**, J. Paetzold, A. Roesner, R. Klar: *Biomedical Data Mining in Clinical Routine: Expanding the Impact of Hospital Information Systems*. Accepted for Medinfo 2007 in Brisbane, Australia.

P. Daumke, K. Markó, M. Poprat, S. Schulz, R. Klar: *Biomedical Information Retrieval Across Languages*. Informatics for Health and Social Care, 32(2):131-147.

2006

P. Daumke: *Ein Framework zur automatischen Indexierung von Volltexten und dessen Anwendung im Information Retrieval*. Bachelorarbeit in Informatik an der FernUniversität in Hagen. 2006

K. Markó, **P. Daumke**, J. Paetzold, A. Zaiß: *ICD-10 Kodierung mit dem MorphoSaurus-System*. In M. Löffler, A. Winter (eds.): Tagungsband der 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS '06), Leipzig, Germany. 2006: 299-300.

K. Markó, **P. Daumke**, U. Hahn: *Cross-Lingual Alignment of Biomedical Acronyms and their Expansions*. Proceedings of the XX International Congress of the European Federation for Medical Informatics (MIE '06), Maastricht, Netherlands. 2006:

857-862.

P. Daumke, K. Markó, S. Schulz: *Morphoogle Eine multilinguale Suchmaschine für das WWW*. In Klaus Haasis, Armin Heinzl Dieter Klump (Eds.), Aktuelle Trends in der Softwareforschung. 2006:133-142.

S. Schulz, K. Markó, **P. Daumke**, U. Hahn, S. Hanser, P. Nohama, R. Andrade, E. Pacheco, M. Romacker: *Semantic Atomicity and Multilinguality in the Medical Domain: Design Considerations for the MorphoSaurus Subword Lexicon*. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06), Genua, Italy. 2006: 1684-1687.

P. Daumke, S. Schulz, K. Markó: *Subword Approach For Acquiring and Cross-Linking Multilingual Specialized Lexicons*. 5th International Conference on Language Resources and Evaluation (LREC '06): Workshop on Acquiring and Representing Multilingual, Specialized Lexicons, Genua, Italy. 2006.

2005

P. Daumke, S. Schulz, K. Markó: *Searching Multilingual Medical Content in the Web*. Technology and Health Care. Volume 13, Number 5. 2005.

U. Hahn, **P. Daumke**, S. Schulz, K. Markó: *Cross-Language Mining for Acronyms and their Completions from the Web*. Proceedings of the 8th International Conference on Discovery Science (DS '05), Singapore. 2005.

S. Schulz Stefan, **P. Daumke**, B. Smith, U. Hahn: *How to Distinguish Parthood from Location in Bioontologies*. Proceedings of AMIA Symposium 2005, Washington D.C.. 2005: 669-673.

P. Daumke, S. Schulz, K. Markó: *Morphoogle - Eine medizinische CLIR Schnittstelle zu einer Web-Suchmaschine*. In: R. Klar, W. Köpcke, K. Kuhn, H. Lax, A. Zaiß (eds.): Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS '05), Freiburg, Germany. 2005: 246-248.

2004

K. Markó, U. Hahn, S. Schulz, **P. Daumke**, P. Nohama: *Interlingual Indexing across Different Languages*. Proceedings of RIAO'04, 7th International Conference "Recherche d'Information Assistée par Ordinateur"(RIAO'04), Avignon, France. 2004: 82-99.

2003

K. Markó, **P. Daumke**, S. Schulz and U. Hahn: *Cross-language MeSH Indexing Using Morpho-Semantic Normalization*. Proceedings of the 2003 American Medical Informatics Association Symposium (AMIA '03), Washington D.C., 2003: 425-429.

P. Daumke, K. Markó, S. Schulz, J. Wermter: Automatische MeSH-Indexierung auf der Basis morphosemantischer Normalisierung. Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 2003; 34 (3): 225-228 (GMDS '03).

J. Wermter, S. Schulz, U. Hahn, K. Markó, **P. Daumke**: *Ein sprachanalytisches Verfahren zur Unterstützung von Laien als Anfragesteller beim medizinischen Dokumenten-Retrieval*. Informatik, Biometrie und Epidemiologie in Medizin und Biologie, 2003; 34 (3): 315-317 (GMDS '03).

Poster

P. Daumke, J. Paetzold, K. Markó: *MorphoSaurus in ImageCLEF 2006: The effect of subwords on biomedical IR*. 7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006.

P. Daumke , K. Markó, M. Poprat, S. Schulz: *Multilingual Biomedical Dictionary*. Proceedings of AMIA Symposium 2005, Washington D.C., 2005: 933.

P. Daumke , S. Schulz, K. Markó: *A CLIR Interface to a Web Search Engine*. Proceedings of AMIA Symposium 2005, Washington D.C., 2005: 934.

P. Daumke , K. Markó, M. Poprat, S. Schulz: *Multilinguales Medizinisches Wörterbuch*. In: R. Klar, W. Köpcke, K. Kuhn, H. Lax, A. Zaiß (eds.): Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS '05), Freiburg, Germany. 2005.

Software-Demonstrationen

P. Daumke, S. Schulz, K. Markó: *A CLIR Interface to a Web Search Engine*. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, August 15-19, 2005, Salvador, Brasil.

K. Markó, **P. Daumke**, S. Schulz: *MorphoSaurus – Multilinguale semantische Indexierung medizinischer Dokumente*. In: R. Klar, W. Köpcke, K. Kuhn, H. Lax, A. Zaiß (eds.): Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS '05), Freiburg, Germany. 2005.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Listings

Kapitel 1

Einführung

1.1 Die Flut textueller Informationen in der Medizin

In der Medizin werden in sehr großem und enorm wachsenden Umfang textuelle Informationen sowohl in der Wissenschaft als auch in Klinik und Praxis erzeugt und benötigt. [Price1963] errechnete, dass sich seit den ersten wissenschaftlichen Zeitschriften im 16. Jahrhundert die wissenschaftliche Literatur alle 15 Jahre verdoppelt. In MEDLINE, der größten verfügbaren Datenbank für medizinische Fachliteratur, wurden allein im Jahr 2006 718.963 Artikel neu hinzugefügt, das sind fast 2.000 Artikel pro Tag. Dementsprechend liegt die jährliche Wachstumsrate an neuen Artikeln in MEDLINE bei ca. 4% mit steigender Tendenz, wie Abbildung 1.1 verdeutlicht. Die US National Library of Medicine (NLM) bietet mit PubMed einen kostenfreien und komfortablen Zugang zur MEDLINE-Datenbank an[1]. In den letzten zwei Jahren wurden monatlich ca. 70 Mio. Anfragen an PubMed gestellt.

Für die Medizin hat diese Entwicklung bedeutende Konsequenzen. Die Medizin wird in immer mehr Unterdisziplinen aufgeteilt, weil Ärzte immer weniger in der Lage sind, das gesamte relevante Wissen eines Faches zu beherrschen. Die steigende Spezialisierung und das zunehmende Wissen führen zu immer aufwändigeren medizinischen Behandlungen mit teuren Medizinprodukten und Pharmazeutika. Diese sind maßgeblich für steigende Kosten im Gesundheitswesen verantwortlich [Ulrich2000].

Auch auf die tägliche Arbeit des Arztes in Klinik und Praxis wirkt sich die wissenschaftliche Informationsflut aus. So ist die Informationsbeschaffung für Ärzte mit hohem Zeitaufwand verbunden. 36% der Ärzte in Deutschland benötigen sowohl für die Recherche als auch für die Durchsicht der Informationen jeweils zwischen drei

[1] Zugriff über `http://www.pubmed.org`, eingesehen im Februar 2007

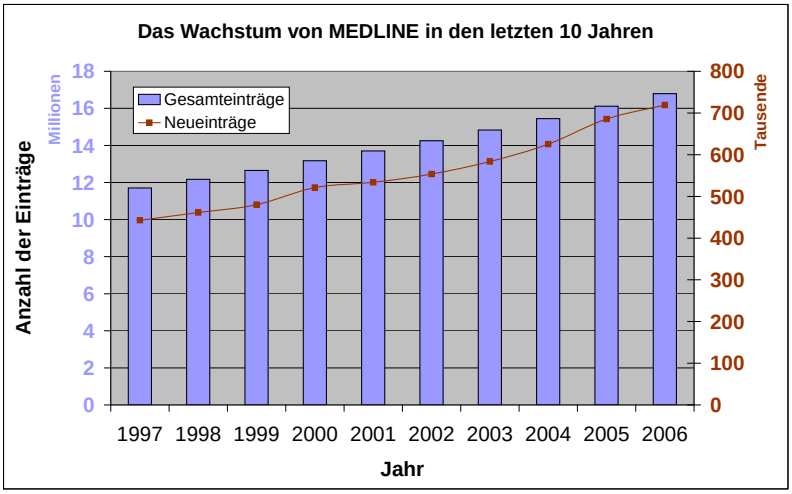

Abbildung 1.1: Grafische Übersicht über den Wachstum von MEDLINE in den letzten 10 Jahren.

und sechs Stunden in der Woche, weitere 40% sind deutlich länger als sechs Stunden pro Woche mit Informationsbeschaffung beschäftigt [Koch & Kaltenborn2005b]. Der hohe Zeitaufwand für die Informationssuche und fehlende Informationen wirken sich negativ auf die Patientenversorgung aus. Kurzfristig äußert sich dies in längeren Wartezeiten und potentiellen Mehrfachuntersuchungen. Langfristig führt die langsame Verbreitung des medizinischen Wissens zu fehlerhaften oder nicht zeitgemäßen Behandlungen. [Antman et al.1992] konnten zeigen, dass die Informationen in Lehrbüchern, Reviews und praktischen Empfehlungen weit hinter dem aktuellen Wissen zurück bleibt. Als Folge hiervon werden nicht effektive Behandlungen durchgeführt oder aktuelle Therapieempfehlungen bleiben unberücksichtigt.

Neben der Informationssuche von genuinem textuellen Wissen in der Medizin gilt es aber auch die textlichen Informationen über einen Patienten besser als bisher für die ärztliche Arbeit zu erschließen. So werden zum Beispiel am Freiburger Universitätsklinikum jährlich 600.000 Befunde, Arztbriefe, OP-Berichte, etc. erstellt, die für die individuelle Patientenversorgung eine viel höhere Relevanz als ICD- oder OPS-Notationen besitzen, jedoch viel schlechter automatisch erschließbar sind.

Der Einsatz innovativer Technologien zur Informationssuche und -beschaffung ist somit eine notwendige Voraussetzung, um die Herausforderungen des modernen Gesundheitswesens in Wissenschaft und Praxis zu bewältigen. Durch diese Technologien können Leistungserbringer im Gesundheitswesen schnell mit relevanten Informationen versorgt werden. Dies ist nicht nur mit enormen ökonomischen Einsparungen verbunden, sondern es stellt auch einen bedeutenden Faktor zur Verbesserung der Behandlungsqualität dar und stellt den Ärzten mehr Zeit für die wesentlichen

Aspekte der Patientenversorgung zur Verfügung. Die derzeitigen Technologien zur Informationssuche werden jedoch als zu komplex und schwer bedienbar eingestuft [Koch & Kaltenborn2005b]. Ein Grund für die Komplexität dieser Technologien sind die vielen sprachlichen Variationen, die besonders in der Sprache der Medizin existieren und für die bisher keine befriedigende rechnerbasierte Lösung gefunden wurde. Diese sprachlichen Herausforderungen werden im Folgenden näher vorgestellt.

1.2 Die medizinische Sprache - Herausforderungen an Retrieval-Systeme

Das Wiederauffinden von Informationen wird gemeinhin als Information Retrieval (IR) bezeichnet, ein System zum Wiederauffinden der Informationen dementsprechend als IR-System. Medizinische IR-Systeme stellen praktische Lösungen für das Gesundheitswesen bereit, so dass Ärzte, Forscher, Angestellte im Gesundheitswesen und nicht zuletzt Patienten gewünschte Informationen aus großen gesundheitsrelevanten Datenmengen herausfinden können. Hierzu gehören Artikel in wissenschaftlichen Zeitschriften, Patienteninformationen aus Gesundheitsportalen, Fachinformationen von Arzneimitteln oder auch Arztbriefe in Krankenhausinformationssystemen.

Eine verbreitete Annahme im Information Retrieval ist, textuelle Dokumente als ein *"bag of words"* (engl. *"Sack voller Wörter"*) anzusehen, in dem die Wörter ungeordnet enthalten sind. Diese Wortmenge kann nach gewissen Kriterien angeordnet werden, um möglichst schnell auf sie zugreifen zu können. Diese Anordnung wird als *Index* bezeichnet. Die *Benutzeranfrage*, die ein Benutzer an ein IR-System stellt, ist ein zweites *"bag of words"*. Die sogenannte "naïve keyword retrieval"-Hypothese [Arampatzis et al.2000] besagt, dass ein Dokument in irgendeiner Form relevant für eine Anfrage ist, wenn in beiden Wortmengen gemeinsame Wörter auftreten. Je größer die Zahl der Übereinstimmung zwischen Anfrage und Dokument ist, desto größer ist nach dieser Hypothese die Relevanz des Dokumentes.

Aus inhaltlicher Sicht stellt die wortzentrierte Sicht auf Dokumente, die in zahlreichen konventionellen IR-Systemen angewendet wird, jedoch nicht immer die richtige Granularität dar [McCray1998]. Konventionelle IR-Systeme basieren auf einfachem Zeichenkettenabgleich zwischen Wörtern, was beim Suchprozess zu einigen Problemen führt. Dies verdeutlicht das folgende Beispiel: Ein Benutzer, der nach *"Blinddarmentzündung"* sucht, wird keine Dokumente finden, die von einer *"Entzündung des Blinddarms"* handeln. Ebenso wenig findet er Dokumente über

"Blinddarmentzündungen, Appendicitis, Appendizitis" oder *"Entzündung der Appendix"*. Grund dafür sind die zahlreichen sprachlichen Variationen, die in einer natürlichen Sprache existieren und die zu dem sogenannten *Vokabularproblem* (*vocabulary problem*) der natürlichen Sprache führen [Furnas et al.1987].

Innovative Retrievaltechnologien stehen vor der Herausforderung, die Besonderheiten der Sprache, in denen die Anfragen und Dokumente verfasst sind, zu berücksichtigen [Strzalkowski et al.1999, Smeaton1999]. Dies betrifft zum einen allgemeine linguistische Phänomene der betrachteten Sprache unabhängig von der Domäne, in der das IR-System eingesetzt wird. Zum anderen sollte ein innovatives IR-System natürlich auch die sprachlichen Besonderheiten der jeweiligen Domäne berücksichtigen, die zusätzliche Anforderungen an das IR-System stellt. Im Folgenden werden zunächst die verschiedenen Ebenen sprachlicher Variationen erörtert, die für das Information Retrieval relevant sind. Anschließend werden die Besonderheiten der medizinischen Sprache näher beschrieben.

1.2.1 Linguistische Variationen in der natürlichen Sprache

Als sprachliche Variationen werden die verschiedenen Möglichkeiten bezeichnet, in denen ein Konzept im Kontext der natürlichen Sprache ausgedrückt werden kann. [Arampatzis et al.2000] unterscheidet, angelehnt an die verschiedenen Ebenen zur Beschreibung einer Grammatik, vier Ebenen sprachlicher Variationen, die für IR-Systeme relevant sind:

- **Morphologische Variationen** beschreiben die verschiedenen Möglichkeiten, in denen ein Wort in der natürlichen Sprache auftreten kann.
- **Syntaktische Variationen** behandeln die sprachlichen Varianten auf Mehrwort- oder Phrasenebene.
- **Lexikalische Variationen** handeln von den verschiedenen Möglichkeiten, ein Konzept zu benennen (Synonymie).
- **Semantische Variationen** schließlich beschreiben den Sachverhalt, dass ein Wort verschiedene Bedeutungen einnehmen kann (Polysemie, Homonymie).

Da lexikalische und semantische Variationen untrennbar miteinander verbunden sind, werden sie oft unter der Bezeichnung *lexiko-semantische Variation* zusammengefasst. Der Autor schließt sich dieser Darstellung an (siehe Abbildung 1.2).

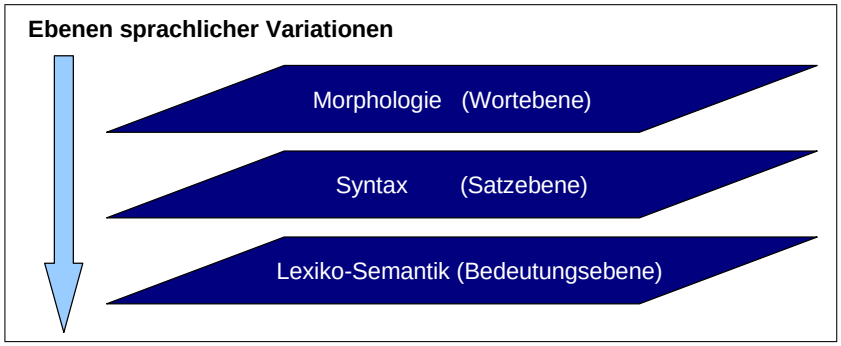

Abbildung 1.2: Übersicht über die verschiedenen Ebenen sprachlicher Variation.

Morphologische Variationen

Die Grundeinheit der Morphologie ist das *Morphem*, das als kleinste bedeutungstragende Einheit einer Sprache definiert ist [Helbig1998]. Ein *freies Morphem* kann als eigenes Wort in einem Satz stehen, ein *gebundenes Morphem* ist immer an ein weiteres Morphem gekoppelt und kann nicht als selbständiges Wort fungieren. Morphologische Variationen beschäftigen sich mit Variationen einzelner Wörter. Diese Variationen entstehen durch verschiedene Verfahren, die bei der Wortbildung von Wörtern zu beobachten sind. Dazu gehören Flexion, Derivation, Komposition und Kürzung. Einen ausführlichen Überblick über morphologische Typologie aus Sicht des Information Retrievals gibt [Pirkola2001].

Flexion bezeichnet die Bildung der grammatischen Wortformen bei flektierbaren Wörtern, oder einfacher ausgedrückt, die Umwandlung von dem, was im Lexikon steht, zu dem, was im Satz vorkommt. Flexion hat keinen Einfluss auf die Wortart des Wortes und wenig Einfluss auf seine Bedeutung. Typische Veränderungen betreffen Pluralbildung bei Nomen (*"Fraktur, Frakturen"*), Komparativbildung bei Adjektiven (*"groß, größer"*) sowie die Bildung der Vergangenheitsformen oder Verlaufsformen von Verben (*"operieren, operiert"*).

Derivation ist die Bildung neuer Wörter aus Bestehendem durch die Kombination von freiem mit gebundenem Morphem. Durch Derivation kann die Wortart eines Wortes verändert werden, wie beispielsweise bei *"operieren"* (Verb) und *"Operation"* (Nomen). Die inhaltliche Bedeutung eines Wortes kann sich durch Derivation in unterschiedlichem Maße ändern.

Komposition ist ein weiterer morphologischer Prozess auf Wortebene und bezeichnet die Bildung von Wörtern aus mindestens zwei freien Morphemen oder Wörtern. Beispiele aus der Medizin sind *"Magenschleimhautentzündung"*, *"Gesundheitsvorsorgeuntersuchung"* oder *"Nebennierenrindeninsuffizienz"*.

Viele indogermanische Sprachen gelten als stark flektierende und agglutinierende Sprachen mit einer Vielzahl von Flexionsendungen und überwiegend mehrsilbigen Wörtern. Dazu gehören Sprachen wie Deutsch, Schwedisch, Finnisch oder Holländisch. [Monz & de Rijke2001] fanden in deutschen und holländischen Zeitungstexten jeweils bis zu 25% Komposita. In anderen Sprachen wie dem Englischen treten weniger Wortbildungsphänomen auf und die Anzahl der Komposita ist geringer. [Schulz & Hahn2000] konnten anhand des englischen ICD jedoch zeigen, dass gerade in der medizinischen Terminologie auch im Englischen zahlreiche Wortkomposita bestehen.

Kürzungen sind ebenfalls eine Form morphologischer Variation. Bei den Kürzungen werden *Abkürzungen* und *Akronyme* unterschieden:

- bei der *Abkürzung* werden die Anfangsbuchstaben der einzelnen Morpheme, aus denen sich das Wort zusammensetzt, aneinander gereiht (*Hämoglobin, Hb*) oder es wird Wortmaterial gelöscht, um ein weniger kompliziertes Wort zu erstellen (*Universität, Uni*).

- das *Akronym* reiht ebenfalls die Anfangsbuchstaben der einzelnen Morpheme aneinander, wobei jedoch ein neues phonetisches Wort entsteht (*Acquired Immunodeficiency Syndrome, AIDS*).

In vielen Anwendungen der natürlichen Sprachverarbeitung wirken sich morphologische Variationen negativ auf deren Funktionalität und Performanz aus. Im Information Retrieval ist dies in besonderem Maße gegeben: Ein IR-System sollte bei der Suche nach *"Magenschleimhautentzündung"* auch Dokumente finden, in denen von einer *"entzündeten Magenschleimhaut"* die Rede ist. Die Herausforderung für IR-Systeme auf morphologischer Ebene liegt also darin, von den morphologischen Variationen zu abstrahieren und die verwendeten Dokumentenkollektionen entsprechend zu normalisieren.

Syntaktische Variationen

Syntaktische Variationen beschäftigen sich mit sprachlichen Variationen oberhalb der Wortebene, also auf Ebene von sogenannten *Syntagmen*. Der Begriff *Syntagma* beschreibt eine grammatikalisch zusammengehörige Wort- bzw. Elementgruppe, wozu Mehrwortausdrücke oder auch ganze Sätze zählen. Mehrwortausdrücke können verschiedener Natur sein. Zum einen gehören hierzu sogenannte lexikalische Einheiten (beispielsweise *"Rotes Kreuz"*), die aus mehreren Wörtern aufgebaut sind, wobei die Elemente für sich genommen nicht den Sinn erkennen lassen. Andere Beispiele

sind *"Multiple Sklerose"* und *"Yellow Fever"*. Zum anderen zählen zu den Mehrwortausdrücken auch sogenannte *Phrasen*, also Satzteile, die nur geschlossen im Satz verschoben werden können. Häufig bestehen Phrasen aus einem *Nomen* (dem *Kopf*) und einem Adjektiv (dem *Modifikator*) (bspw. *"mitochondriale Membran"*), welche der Gruppe der *Nominalphrasen* zugeordnet wird. Andere Phrasen sind *Verbal-, Präpositional-, Adjektiv-* und *Adverbphrasen*. Schließlich gehören zu den Mehrwortausdrücken auch feste Redewendungen oder Phraseologismen wie *"ins Gras beißen"* oder *"auf dem Holzweg sein"*.

[Daille et al.2000] und [Jacquemin & Tzoukermann1999] unterscheiden drei Arten syntaktischer Variation:

- **Variationen durch Einfügung oder Auslassung (Substitution)**: betrifft Wörter innerhalb eines Mehrwortausdruckes, die nicht nur grammatische Bedeutung haben. Beispielsweise ist der Ausdruck *"zweidimensionale tryptische Peptidanalyse"* mit *"zweidimensionaler Analyse"* assoziiert.

- **Koordinationsvariationen (Koordination)**: betrifft koordinierte (aneinander gereihte) Wörter oder Phrasen. So betrifft der Ausdruck *"Systolischer und diastolischer Blutdruck"* sowohl den *systolischen* Blutdruck als auch den *diastolischen* Blutdruck.

- **Permutationsvariationen (Permutation)**: betrifft alle Wörter innerhalb einer Phrase, die um ein bestimmtes Schlüsselwort herum vertauscht werden können. Beispielsweise ist *"diseases of the urinary tract"* (*"Erkrankungen des Harnapparates"*) eine syntaktische Variante von *"urinary tract disease"*.

Die Herausforderungen für Retrievalsysteme sind dementsprechend, syntaktische Varianten zu erkennen und richtig zu interpretieren. Die Anfrage *"Fraktur des Femurs"* sollte auch Dokumente finden, in denen die Phrase *"[. . .] ist zusätzlich das Femur frakturiert."* erscheint (Permutation). Bei der Anfrage *"zweidimensionale tryptische Peptidanalyse"* sollte ein Retrievalsystem zwar Dokumente über *"zweidimensionale Peptidanalysen"* finden, nicht jedoch Dokumente über *"zweidimensionale Peptide"* (Substitution).

Feststehende Mehrwortbegriffe stellen einen Sonderfall für Retrievalsysteme dar, wie die folgenden Beispiele verdeutlichen:

- *"Blood Bank"* (*"Blutbank"*) ist keine aus Blut hergestellte Bank, sondern eine Abteilung im Krankenhaus, die für die Aufbewahrung von Blut zuständig ist.

- *"Athlete's Foot"* bezeichnet meist nicht den *"Fuß des Athleten"*, sondern ist die englische Bezeichnung für das deutsche Wort *"Fußpilz"*.

Dementsprechend sollten Ausdrücke wie *"The bank was full of blood"* im ersten Beispiel und *"the foot of the athlete was swollen"* im zweiten Beispiel nicht als relevant bewertet werden.

Lexiko-Semantische Variationen

Da lexikalische und semantische Variationen untrennbar miteinander verbunden sind, werden sie in diesem Abschnitt gemeinsam behandelt. Unter dem Begriff *Lexikalische Variation* wird jegliche Form der *Synonymie* zusammengefasst, also die semantische Gleichheit oder Ähnlichkeit verschiedener Wörter. Synonyme Wortpaare sind beispielsweise *"Hypertonie/Bluthochdruck"* oder *"Pankreas/Bauchspeicheldrüse"*. Synonymie betrifft sowohl einzelne Wörter als auch Mehrwortausdrücke. Ein Beispiel für Synonymie auf Mehrwortebene stellen die Synonyme für *"Sarkoidose"* dar: *"Besnier-Boeck-Schaumann-Krankheit", "Lymphogranulomatosis benigna", "Sarkoidose"*.

Zu den *Semantischen Variationen* werden die *Polysemie* und die *Homonymie* gezählt, also das Phänomen, dass einzelne Wörter verschiedene Bedeutungen besitzen, wenn sie in unterschiedlichem Kontext auftreten. *Polyseme* müssen miteinander verbunden oder auf eine gemeinsame etymologische Wurzel zurückzuführen sein. Beispielsweise bezeichnet *"Atlas"* den ersten Halswirbel, aber auch ein Gebirge in Nordwest-Afrika, einen Mond des Saturns, einen Stern in den Pleiaden, ein kartografisches Werk, ein hochglänzendes Seidengewebe und einiges mehr[2]. Alle Bezeichnungen stammen ursprünglich von dem Titanen der griechischen Mythologie ab, der von Zeus auf ewig dazu verbannt wurde, auf seinen Schultern das Himmelsgewölbe zu tragen. Bei der *Homonymie* dagegen stammen die Homonyme von unterschiedlichen Wurzeln ab. Beispielsweise ist *ATLAS* auch ein Akronym für *A Toroidal LHC AparatuS*, ein Teilchenbeschleuniger in CERN, und stammt daher nicht von der griechischen Sagenfigur ab.

Neben der Synonymie können darüber hinaus hierarchische Beziehungen zwischen verschiedenen Wörtern bestehen, die für IR-Systeme eine Rolle spielen können. Die zwei bekanntesten sind

- *Hyponymie / Hyperonymie* (Die *"Is-A Relation"*): Diese Relationen definieren Ober- und Unterbegriffe von Wörtern in einem Thesaurus. Beispielsweise ist

[2]siehe Wikipedia - Die freie Enzyklopädie, http://de.wikipedia.org/wiki/Atlas, eingesehen im Februar 2007.

Penizillin ein Unterbegriff von *Antibiotikum*. Sind solche Relationen zwischen Wörtern oder Mehrwortausdrücken definiert, ist es beispielsweise möglich, bei der Suche nach *"Nebenwirkungen von Antibiotika"* auch alle Dokumente zu ermitteln, die von *"Nebenwirkungen von Penizillinen"* handeln.

- *Meronymie / Holonymie* (Die *"Part-Of Relation"*): Teil-Ganzes-Beziehungen spielen zum Beispiel in der medizinischen Anatomie eine besondere Rolle. So ist der *"Magen"* Teil des *"Gastrointestinalsystems"*, die *"Magenschleimhaut"* ist Teil des *"Magens"*. Ein intelligentes IR-System kann bei der Anfrage nach *"Ulzera im Gastrointestinaltrakt"* auch alle Dokumente anzeigen, die von *"Ulzera der Magenschleimhaut"* handeln.

Eine besondere Herausforderung für IR-Systeme bezüglich lexiko-semantischer Variation stellt daher die Identifikation relevanter Synonyme einer Anfrage mit dem Ziel dar, Dokumente anzuzeigen, die entweder Wörter der Originalanfrage oder relevante Synonyme enthalten. In ausgewählten Fällen könnte es darüber hinaus sinnvoll sein, die anderen "-nymie"-Beziehungen zu verwenden, wie in obigen Beispielen zur Hyponymie und Meronymie gezeigt. Außerdem sollte ein IR-System Mehrdeutigkeiten erkennen und versuchen, diese kontextbezogen aufzulösen. Unter Berücksichtigung der Mehrdeutigkeit des Wortes *"Bruch"*, welches zum einen die *"Hernie" ("Weichteilbruch")*, zum anderen die *"Fraktur" ("Knochenbruch")* bezeichnet, sollte bei der Anfrage nach *"Leistenbruch"* möglichst nur diejenigen Dokumente als relevant bewertet werden, die von *"Leistenhernien"* handeln, dagegen aber keine Dokumente, in denen der Ausdruck *"Frakturen im Bereich der Leiste"* erscheint.

1.2.2 Linguistische Besonderheiten der medizinischen Sprache

Die medizinische Sprache ist wie alle Fachsprachen charakterisiert durch ihre enorme Fülle an domänenspezifischer Terminologie. [Fluck1996] beziffert den Gesamtwortschatz der deutschen medizinischen Terminologie, des "Thesaurus linguae medicinae", im Jahr 1996 auf circa 500.000 Termini. Das Unified Medical Language System (UMLS) [Umls2005a], ein Metathesaurus, der über 100 internationale medizinische Terminologien integriert, enthält über 1,3 Millionen Konzepte und 6,4 Millionen Konzeptbezeichner in verschiedenen Sprachen.

Die medizinische Sprache ist geprägt aus einem Gemisch aus Wörtern gepaart mit griechischen und lateinischen Wurzeln, ein Sachverhalt, der als *neoklassische Wort-*

bildung [McCray et al.1988] bezeichnet wird. Bei der *neoklassischen Wortbildung* werden Wortbildungselemente verwendet, die keine eigenständigen Wörter sind. Diese Elemente stammen aus klassischen Sprachen und sind meist griechischen und in geringerem Maße lateinischen Ursprungs. Sie können sich miteinander oder mit anderen Wortteilen zu neuen Wörtern zusammensetzen. Da die mit ihnen gebildeten Wörter oft moderne Bildungen nach klassischem Muster sind, wird das Wortbildungsverfahren "neoklassisch" genannt. Typische Beispiele aus der deutschsprachigen Medizin sind *"Ösophagoduodenoskopie", "Adenotonsillektomie"* oder *"Acromioklavikulargelenk"*.

Der große Umfang des medizinischen Wortschatzes lässt den starken Benennungsbedarf des Faches deutlich werden. Durch den ständigen Fortschritt aus den verschiedenen Fachgebieten der Medizin entstehen immer wieder neue synonyme Bezeichnungen [Lüking1994], wie am Beispiel der *"Fallot-Tetralogie"* deutlich wird: *"Fallot-Syndrom", "Morbus caerleus", "Corvisart-Komplex", "Blue baby", "Maladie de Fallot", "Maladie bleu", "Cyanose cardiaque", "Fallot"*. Auffällig ist die hohe Zahl an orthografischen Varianten besonders in der deutschen medizinischen Terminologie [Schulz et al.2002], wie an den verschiedenen Schreibweisen für den lateinischen Ausdruck von *"Blinddarm"* deutlich wird: *"Zäkum", "Cäkum", "Zaekum", "Caekum", "Zaecum", "Caecum"*. Auch die Verwendung von Eigennamen ist in der Medizin durchaus weit verbreitet. So bezeichnet der Ausdruck *"Bruch"* nicht nur eine *"Hernie"* und eine *"Fraktur"*, sondern er wird auch für die sogenannte *"Bruch-Membran"* verwendet, einer Grenzschicht im Bereich der Netzhaut, die nach ihrem Entdecker Karl Wilhelm Bruch (1819-1884) benannt wurde. Im weiteren Sinne können zu den Eigennamen auch die Produktnamen der Pharmazeutika gezählt werden, die je nach Kontext synonym zu dem entsprechenden Wirkstoff verwendet werden (*"Aspirin", "ASS", "Acetylsalicylsäure"*).

Die Medizin weist mit dem Selbstverständnis als Wissenschaft von der Erkenntnis und der Behandlung menschlicher Krankheiten eine ausgeprägte horizontale und vertikale Schichtung auf [Lüking1994]. Aus horizontaler Schicht gliedert sich die Medizin in ihre verschiedenen Teildisziplinen wie Anatomie, Physiologie, Innere Medizin, Chirurgie und bringt nicht eben geringe interdisziplinäre Kommunikationsprobleme mit sich. Durch die horizontale Schichtung ergeben sich in der Medizin sogenannte Fachumgangssprachen mit individueller Note, eine Art medizinischer Slang. Beispielsweise hat sich zur Beschreibung pathologischer Befunde eine teils blumig-metaphorische Sprache entwickelt, die im Kontrast zu der sonst sehr exakt, sachlich und eindeutig beschreibenden medizinischen Terminologie steht, wie [Heckl2006] am Beispiel der *"kleinhühnerfaustgroßen Schwellung"* ironisch pointiert. Im klinischen

Bereich hat sich eine aus der täglichen Praxis entstandene, leicht handhabbare Kurzform herausgebildet. Der Ausdruck "58 yo male c/o pain r leg, o/e NAD" bedeutet "Zur ärztlichen Untersuchung kam ein 58 jähriger Patient, der über Schmerzen im rechten Bein klagte; die anschließende Untersuchung blieb unauffällig." ("58 year old patient with complaint of pain in the right leg; on examination nothing abnormal detected").

Die vertikale Schichtung der Medizin umfasst drei Ebenen: die erste Ebene ist die Wissenschaftsebene im engeren Sinne, auf der über medizinische Erkenntnisse kommuniziert wird, die zweite Ebene beschreibt die Praxisebene, auf der zwischen Ärzten und medizinischem Fachpersonal kommuniziert wird und die dritte Ebene schließlich ist die Behandlungsebene zwischen Ärzten und Patienten [Roelcke2005]. Diese vertikale Schichtung spiegelt sich auch in der Verwendung des Englischen als international verbreitete Fachsprache wider. So wird auf der Wissenschaftsebene nahezu ausschließlich auf Englisch kommuniziert, und Veröffentlichungen in der Landessprache wird ein englisches Abstract vorangestellt. Aufgrund des stetigen internationalen Fortschritts werden auch auf der zweiten Ebene immer mehr Anglizismen wie *"Zerebraler Blood Flow", "Downregulation"* oder *"Compliance des Patienten"* eingeführt [Heckl2006]. Auf der dritten Ebene ist die Zahl der Anglizismen hingegen gering.

Zwischen den einzelnen Schichten entstehen naturgemäß kommunikative Probleme, da die medizinische Fachterminologie für den medizinischen Laien unverständlich ist. "Der lebensnahen, aber mehrdeutigen Alltagssprache des Patienten steht die zunehmende Begriffsgenauigkeit der medizinischen Fachsprache des Arztes gegenüber, deren Wörter als Wörter einer Wissenschaftssprache zwar relativ eindeutig und damit für das Verständnis unter Ärzten unverzichtbar, aber doch auch oft zu arm sind, um die Komplexität und Vieldeutigkeit konkreter Phänomene der Wirklichkeit immer angemessen erfassen zu können" [Koch & Kaltenborn2005a].

1.3 Ansätze zum Umgang mit linguistischen Herausforderungen

Im vorherigen Abschnitt wurden die Herausforderungen der medizinischen Dokumentenrecherche beschrieben, die sich durch die Variationen der natürlichen Sprache ergeben. Diesen Herausforderungen steht eine Vielzahl von Lösungsansätzen gegenüber, die verschiedene Techniken der natürlichen Sprachverarbeitung verwenden und dabei sprachliche Ressourcen wie Lexika und Thesauri einsetzen. Das Ziel, ein vollständiges und eindeutiges Verständnis von Sprache zu erreichen, hat sich dabei

als schwierig herausgestellt und der Erfolg linguistischer Methoden im Information Retrieval ist insgesamt moderat [Hersh2002]. Dennoch gibt es einige Techniken, die die Retrievalergebnisse in Experimenten signifikant steigern und sich mittlerweile als Bestandteile in IR-Systemen etabliert haben. Diese sollen in den folgenden Abschnitten genauer beschrieben werden.

1.3.1 Morphologische Ansätze

Stammformbildung

Die englische Sprache hat recht einfache Wortflexionsregeln, was dazu führte, dass theoretische Forschung auf dem Gebiet der Morphologie zunächst auf wenig Interesse bei IR-Forschern stieß [Airio2006]. Erst durch das wachsende Interesse an der Dokumentenrecherche in nicht-englischen Sprachen, gerade auch mit dem Aufkommen der mehrsprachigen Suche (*Cross-Language Information Retrieval*), haben sich Forscher mit den Phänomenen in morphologisch komplexeren Sprachen auseinandergesetzt.

Das bedeutsamste linguistische Verfahren für das Information Retrieval ist die sogenannte *Stammformbildung* (*Stemming*). Damit wird die Umformung verschiedener morphologischer Variationen zu einem gemeinsamen Wortstamm bezeichnet. Typischerweise verwenden Algorithmen wie der Lovins-Stemmer [Lovins1968] oder der Porter-Stemmer [Porter1980] bei der Stammformbildung keine Wörterbuchinformationen, sondern basieren allein auf einer Liste von Verkürzungsregeln sowie einigen Umformungsregeln. Die einfachsten Stemmer wie der Lovins-Stemmer beruhen auf einer schrittweisen Entfernung von Suffixen (Wortendungen) vom Wortende, wobei immer das längstmögliche Suffix entfernt wird und das Wort nur einmal die Menge der Verkürzungsregeln durchläuft. Anschließend finden einige Umformungen statt. Der Porter-Stemmer basiert auf einer Menge von Verkürzungsregeln, die so lange auf ein zu stemmendes Wort angewendet werden, bis dieses eine Minimalanzahl von Silben aufweist. Die Verkürzungsregeln bestehen aus Paaren von Bedingungen und Ableitungen für verschiedene Suffixe. Die Regeln sind in Gruppen zusammengefasst, die nacheinander abgearbeitet werden. Aus jeder Gruppe darf nur eine Regel angewendet werden. Beispiele für verschiedene Verkürzungsregeln sind “sses” → “s”, “ies” → “i” und “s” → “”. Der ursprünglich für die englische Sprache entwickelte Algorithmus kann relativ leicht für andere Sprachen portiert werden und liegt mittlerweile in 14 verschiedenen Sprachen vor[3].

Die Nützlichkeit der Stemming-Verfahren auf das Information Retrieval in verschiedenen Sprachen wurde kontrovers diskutiert ([Harman1991, Krovetz1993]

[3] `http://snowball.tartarus.org/`, eingesehen im Februar 2007

[Hull1996, Kantrowitz et al.2000, Tomlinson2001, Braschler & Ripplinger2004, Tordai & de Rijke2005]). Insbesondere in morphologisch komplexen Sprachen sind solche Verfahren oft nicht ausreichend. Eine Schlüsselrolle für die Qualitätsverbesserung der Stemming-Verfahren scheint in der Anwesenheit lexikalischer Komponenten zu liegen, die lexikalische Grundformen beinhalten, sowie möglicherweise in der Anwesenheit von NLP-Werkzeugen, die bestimmte linguistische Informationen wie Kasus, Genus oder Numerus ergänzen. So konnte [Krovetz1993] zeigen, dass Wörterbuch-basierte Stemmer eine deutliche Verbesserung der Stemming-Ergebnisse erzielen.

Lemmatisierung

Die Bildung der Grundform eines Wortes wird *Lemmatisierung* genannt (*"traf, treffe, treffen, trifft"* → *"treffen"*). Computerbasierte Lemmatisierer verwenden Lexika, in denen lexikalische Variationen von Wörtern eingetragen sind. [Niedermair et al.1984] verwenden ein Morphemlexikon, eine Morphemgrammatik und einen Algorithmus zur automatischen Zerlegung von Wörtern in ihre Wortbestandteile und zur Überführung in das jeweilige Lemma. Durch die Grammatik werden Elemente eines Wortes in Präfixe, Stämme, Derivation- und Flexionsendungen gruppiert.

[Koskenniemi1984] beschreibt ein zweistufiges Modell eines Lemmatisierers, der aus einem Lexikon und einem Regelsystem aufgebaut ist. Die Regeln definieren, wie Affixe an Wörter angefügt werden können. Das Modell ist sprachunabhängig, neue Sprachen können durch sprachspezifische Lexika und Anpassung der Regeln hinzugefügt werden. Aus dieser Arbeit resultieren Lemmatisierer für das Finnische, Schwedische, Englische und Deutsche.

[Lezius et al.1998] beschreiben einen weiteren Ansatz zur Lemmatisierung. In ihrem Morphologiesystem *Morphy* existiert für jede Wortklasse (Substantive, Adjektive, schwache Verben, Eigennamen usw.) eine eigene Datenstruktur. Daher gibt es für jede Wortklasse ein eigenes Unterlexikon und einen eigenen Algorithmus für die Lemmatisierung. Um eine Wortform zu lemmatisieren, wird versucht, durch Abschneiden aller möglichen Endungen auf potentielle Wortstämme zu schließen. Anschließend wird überprüft, ob die Kandidaten im wortklassenspezifischen Grundform-Lexikon enthalten sind.

Lemmatisierer haben den Nachteil, dass sie für verschiedene Wortarten unterschiedliche Lemmata ausgeben, was für das Information Retrieval nicht immer sinnvoll ist. Beispielsweise werden die Wörter *"operierte"* und *"Operationen"* auf die unterschiedlichen Lemmata *"operieren"* und *"Operation"* zurückgeführt. Dies ist beim

IR jedoch nicht gewünscht, da dort versucht wird, unabhängig von der Wortart alle Variationen eines Begriffes der Suchanfrage zu berücksichtigen.

Weder Stemmer noch Lemmatisierer berücksichtigen die Zerlegung zusammengesetzter Wörter, welche jedoch für das Information Retrieval eine besondere Rolle spielt. [Monz & Dorr2005] fanden eine deutliche Verbesserung der Retrievalergebnisse im Deutschen und Holländischen durch die Verwendung von Kompositazerlegung. [Braschler & Ripplinger2004] zeigten, dass Dekomposition im Deutschen positivere Auswirkungen auf die Informationsrecherche hat als Stemming.

Bei der Zerlegung zusammengesetzter Wörter werden verschiedene Techniken eingesetzt. Einige Techniken zerlegen nur diejenigen Komposita, in denen die Konstituenten dieselbe Wortart besitzen (Nomen/Nomen, Adjektiv/Adjektiv) und werden daher als "konservativ" bezeichnet. Andere Verfahren zerlegen Komposita in alle möglichen Wortformen, oftmals unterstützt durch ein zusätzliches Grundformenlexikon. Da diese Verfahren bei der Zerlegung keine Wortarten berücksichtigen, also auch in Pronomen, Präpositionen oder Artikel zerlegen, werden diese Verfahren als "aggressiv" bezeichnet [Braschler & Ripplinger2004].

Linguistisch motivierte Ansätze liegen zwischen konservativen und aggressiven Verfahren und zerlegen Komposita nur in erlaubte Wortarten wie Nomen, Verben und Adjektive. [Chen & Gey2004] schlagen folgenden konkreten Algorithmus für die Zerlegung von Komposita vor, mit dem sie eine Steigerung ihrer Retrievalergebnisse auf deutschen Texten um über 10% erzielen: Voraussetzung ist, dass ein Wörterbuch mit Grundformen (also ohne zusammengesetzte Wörter) vorliegt beziehungsweise erzeugt wird. Dieses enthält Informationen über die Häufigkeit des Auftretens dieser Wörter in typischen Texten der Sprache. Ein Eingabewort wird dann auf Basis dieses Lexikons in alle möglichen Zerlegungen überführt. Unter diesen Zerlegungen werden diejenigen mit der kleinstmöglichen Anzahl von Fragmenten ausgewählt. Anschließend werden für die verbleibenden Zerlegungen anhand der ermittelten Häufigkeiten die wahrscheinlichste Zerlegung ermittelt.

[Monz & de Rijke2001] beschreiben einen Ansatz zur Zerlegung deutscher Nomen-Nomen Komposita. Am Anfang eines Wortes beginnend werden rekursiv mögliche Nomen vom Kompositum entfernt, sofern das restliche Teilwort ebenfalls in Wörter aus einem zur Verfügung stehenden Lexikon zerlegt werden kann.

[Savoy2003] schlagen einen Ansatz für die deutsche Kompositazerlegung vor, der vollständig ohne lexikalische Ressourcen auskommt und auf einer Menge vordefinierter Muster basiert. Beispielsweise werden Wörter, in denen die Zeichenfolge "gss" erscheint (*Betreuungsstelle*), in zwei Wörter mit der Endung "g" (*Betreuung*) und dem Anfangsbuchstaben "s" (*Stelle*) zerlegt.

1.3.2 Syntaktische Ansätze

Konventionelle Suchdienste, die sich dem Paradigma des "naïve keyword retrieval" verschrieben haben, berücksichtigen weder die Wortreihenfolge, in der eine Anfrage formuliert ist, noch die Reihenfolge, in der die Wörter in den Dokumenten vorliegen. Dies ist bereits ein primitiver Ansatz, um syntaktische Variationen zu berücksichtigen. Beispielsweise werden von diesen Verfahren bei der Suche nach *"urinary tract infection"* auch Dokumente gefunden, in denen *"infection of the urinary tract"* enthalten ist. Allerdings berücksichtigen diese Verfahren keine Mehrwortausdrücke, die als lexikalische Atome angesehen werden. So wird bei der Anfrage *"Yellow Fever"* auch das Dokument mit dem Inhalt *"Yellow Sputum and Fever"* als relevant gewertet. Eine einfache Erweiterung besteht darin, Benutzern von Retrievalsystemen die Möglichkeit zu geben, Nähebeziehung zwischen den Anfragewörtern zu definieren. Beispielsweise kann definiert werden, dass zwischen *"Yellow"* und *"Fever"* kein dazwischen liegendes Wort im Dokument erscheinen darf. Eine solche Anfrage wird in IR-Systemen im Allgemeinen in Anführungszeichen gesetzt (*"Yellow Fever"*). Es kann auch nur eine bestimmte Anzahl von Wörtern zwischen den Eingabewörtern zugelassen werden. Wird ein Wort zwischen *"Yellow"* und *"Fever"* als Zwischenwort zugelassen, kann beispielsweise auch der Ausdruck *"Yellow/Dengue Fever"* als relevant beurteilt werden. Eine solche Nähebeziehung kann in einigen Suchmaschinen beispielsweise durch die Syntax *"Yellow Fever"*$\sim$*3* erreicht werden, welche ausdrückt, dass die Worte *"Yellow"* und *"Fever"* in einem Fenster von maximal drei Wörtern gefunden werden müssen.

Aufwändigere Techniken zur Identifikation syntaktisch zusammengehöriger Gruppen verzichten auf Benutzerinteraktion und versuchen, entweder basierend auf statistischen Daten oder basierend auf syntaktischer Analyse zusammengehörige Wortgruppen zu identifizieren und anschließend so umzuformen, dass auch syntaktische Varianten von Benutzeranfragen in Dokumenten gefunden werden [Arampatzis et al.2000]. Bei statistischen Verfahren wird bestimmt, wie häufig bestimmte Wörter gemeinsam auftreten. Syntaktische Verfahren analysieren die verschiedenen Wortarten aus Anfragen und Dokumentenkollektionen und ermitteln syntaktische Muster, die auf Mehrwortausdrücke hindeuten. Dafür werden ausgefeilte NLP-Techniken benötigt. Ob statistische oder syntaktische Verfahren für das Information Retrieval nützlicher sind, bleibt unklar [Mitra et al.1997, Wessel Kraaij1998].

1.3.3 Semantische Ansätze

Erweiterung von Anfragen

Im Umgang mit lexikalischen Variationen im Information Retrieval sind zwei Verfahren von besonderem Interesse, zum einen die Anfrageerweiterung basierend auf den Suchergebnissen (Relevance Feedback) und zum anderen die Anfrageerweiterung basierend auf Wissensstrukturen wie Thesauri [Efthimiadis1996, Baeza-Yates & Ribeiro-Neto1999].

Bei der Thesaurus-basierten Anfrageerweiterung werden Synonyme explizit zu einer Anfrage hinzugefügt. Wenn eine Frage nach *"Bauchspeicheldrüsenentzündung"* erfolgt, könnte die Anfrage erweitert werden zu *"Bauchspeicheldrüsenentzündung OR Pankreatitis OR Entzündung des Pankreas"*. Hierzu werden entweder manuell erstellte Thesauri wie WordNet [Fellbaum1998] oder automatisch erstellte Thesauri [Grefenstette1994, Cooper & Byrd1997] verwendet. Die Anfrageerweiterung kann einerseits in einem interaktiven Prozess zwischen Benutzer und IR-System erfolgen, oder das IR-System erweitert eine Benutzeranfrage automatisch, wobei jedoch der Erfolg der automatischen Anfrageerweiterungen nicht endgültig geklärt ist [Voorhees1994, Smeaton et al.1995]. Die Gefahr bei der Ergänzung von Synonymen besteht darin, dass eine Unschärfe in die Anfragen eingeführt wird, die den positiven Nutzen der Anfrageerweiterung überdeckt.

Relevance Feedback ist ein Verfahren, bei dem die Suchanfrage basierend auf den Suchergebnissen der ersten Anfrage und deren Relevanzbeurteilung neu formuliert wird [Baeza-Yates & Ribeiro-Neto1999]. Die Beurteilung der Relevanz kann dabei vom Benutzer oder vom IR-System erfolgen. Beim *User Relevance Feedback* (*Benutzer-Relevanz-Rückmeldung*) entscheidet der Benutzer über ein Interface, welche Dokumente seiner ursprünglichen Anfrage aus der Antwortmenge relevant sind. Beim sogenannte *Blind Relevance Feedback* (*BRF, auch Pseudo Relevance Feedback*) bestimmt das IR-System anhand festgelegter Kriterien, welche Dokumente als relevant einzustufen sind. Dies können beispielsweise die ersten zehn Dokumente einer Treffermenge sein. Auf Basis dieser Relevanzbeurteilung wird eine neue Anfrage gestellt, in der zum einen die bestehenden Terme ein neues Gewicht erhalten und zum anderen neue als relevant eingestufte Terme zur Anfrage hinzugefügt werden können.

Bereits in den 1970ern führte [Rocchio1971] erste RF-Experimente im SMART Retrieval System durch. Seitdem ist in verschiedenen Ansätzen ([Buckley et al.1994, Xu et al.2004, Robertson et al.2000] die Nützlichkeit von RF im Information Retrieval gezeigt worden. [Carpineto et al.2001] führten umfangreiche Tests zu verschie-

denen Term-Ranking-Funktionen durch, die im *Relevance Feedback* eingesetzt werden. Zu den bekanntesten Ranking-Funktionen zählen der *Robertson-Selection-Value* [Robertson1990] sowie die *Kullback-Leibler-Divergenz* [Kullback & Leibler1951].

Auflösung von Mehrdeutigkeiten

Techniken zum Umgang mit semantischen Variationen von Wörtern werden gewöhnlich unter dem Begriff *Word-Sense-Disambiguation (WSD)* (*"Auflösung von Mehrdeutigkeiten"*) zusammengefasst. Da die Mehrdeutigkeit von Begriffen schon lange als ein Phänomen angesehen wird, welches sich nachteilig auf die Trefferqualität von Retrieval-Systemen auswirkt, sind die Techniken zur *Word-Sense-Disambiguation (WSD)* im Kontext des Information Retrievals intensiv erforscht. Arbeiten hierzu reichen bis in die 1950er zurück [Kaplan1955]. Interessanterweise erzielen diese Techniken nicht immer gute Ergebnisse und zeigen sich bisweilen sogar als nachteilig, gerade wenn die Disambiguierung nicht exakt genug erfolgt [Gale et al.1992]. Mit der Ausnahme von [Schütze & Pedersen1995] haben nur wenige Autoren über signifikante Verbesserungen ihrer Retrieval-Ergebnisse durch Disambiguierung berichtet. Dies bewog Autoren zu der Aussage, dass Disambiguierungswerkzeuge erst ab einer Genauigkeit von 90% [Sanderson1994] beziehungsweise ab einer Genauigkeit, die an menschliche Disambiguierung heranreicht [Gale et al.1992], für das Information Retrieval einen positiven Nutzen darstellen. Einen ausführlichen Überblick über Wortsinndisambiguierung (WSD) geben [Ide & Veronis1998], verschiedene Techniken von WSD im Information Retrieval beschreiben [Sanderson2000, Ide & Véronis1998].

1.4 Überblick über diese Arbeit

In dieser Arbeit wird mit dem MorphoSaurus-System ein neuartiger Ansatz vorgestellt, der Lösungen für viele der linguistischen Herausforderungen im Information Retrieval bietet. Bei diesem Verfahren wird der semantischen Sichtweise auf Dokumente dahingehend Rechnung getragen, dass Texte nicht in einzelne Wörter zerlegt werden, sondern in *inhaltlich atomare* Einheiten. Diese Einheiten werden als *Subwörter* bezeichnet. Der Vorgang, bei dem die Originalwörter in die semantisch atomare Form der Subwörter überführt werden, nennt sich *morpho-semantische Normalisierung*.

Subwörter sind in speziellen, sprachspezifischen *Subwort-Lexika* gespeichert. Beispielsweise sind in dem deutschen Subwort-Lexikon die Subwörter *"appen-*

diz" und *"itis"* enthalten, jedoch nicht das Wort *"appendizitis"*. Zusammengesetzte Wörter wie *"appendizitis"* werden durch die morpho-semantische Normalisierung in die einzelnen Bestandteile *"appendiz"* und *"itis"* aufgetrennt. Es können auch Mehrwort-Einträge wie *"morbus boeck"* in die Subwort-Lexika eingetragen werden. Über einen Thesaurus sind Subwörter gleicher Bedeutung in sprachunabhängige *Äquivalenzklassen* zusammengefasst. Dort sind *"appendiz"* mit *"blinddarm"*, *"itis"* mit *"entzünd"* und *"morbus boeck"* mit *"sarkoidos"* verknüpft. In der Äquivalenzklasse für *"Entzündung"* befindet sich auch das englische Subwort *"inflamm"*, das spanische Subwort *"inflam"* sowie entsprechende Ausdrücke in anderen Sprachen. Somit ist die mehrsprachige Dokumentenrecherche möglich. Die Verwendung von Subwörtern ermöglicht es, selbst große Begriffswelten wie die der Medizin mit einem relativ kleinen Inventar an lexikalischen Einträgen abzudecken. Dies umgeht das Problem der kombinatorischen Explosion von Begriffen, wie dies bei der Erstellung von Vollformen-Synonymlisten in stark flektierenden und agglutinierenden Sprachen üblich ist. Die wesentlichen Konzepte des MorphoSaurus-Systems werden in Kapitel 2 eingeführt und die Realisierung der MorphoSaurus-Konzepte in der medizinischen Domäne beschrieben.

Kapitel 3 beschreibt die einsprachigen IR-Experimente, die mit MorphoSaurus durchgeführt wurden. Als Dokumentenkollektionen kommen das OHSUMED-Korpus [Hersh et al.1994a], die ImageCLEF-Kollektionen von 2006 [Müller et al.2006a] und das GIRT-Korpus [Kluck2004] zum Einsatz. Während OSHUMED und ImageCLEF Kollektionen aus der Domäne der Biomedizin darstellen, ist GIRT ein sozialwissenschaftliches Korpus. Damit soll die Einsatzfähigkeit des MorphoSaurus-Systems in einer anderen Domäne als der Medizin getestet werden.

Kapitel 4 beschäftigt sich mit der Verschlagwortung biomedizinischer Artikel mit Einträgen aus dem MeSH-Thesaurus [MeSH2006]. Die Verschlagwortung von Artikeln dient dazu, von den semantischen und sprachlichen Variationen, die im Information Retrieval vernachlässigt werden können, zu abstrahieren und eine einheitliche Repräsentation der Artikel in Form mehrerer Schlagwörter zu erreichen, so dass relevante Dokumente besser identifiziert werden können. Dabei können bestehende Relationen von Thesauri wie Ober- und Unterbegriffe ausgenutzt werden, um beispielsweise auch Dokumente über *"Penizilline"* zu finden, wenn nach Informationen über *"Antibiotika"* gesucht wird.

Ein Ansatz zur Übersetzung von Wörtern und Mehrwortausdrücken basierend auf dem MorphoSaurus-System wird in Kapitel 5 vorgestellt. Er behandelt insbesondere ein typisches Problem der automatischen Termübersetzung, und

zwar die Übersetzung sogenannter *Out-Of-Vocabulary-Terms*, also die Übersetzung von Ausdrücken, die nicht in konventionellen Lexika enthalten sind. Dieses Übersetzungsverfahren kann zum Beispiel im mehrsprachigen Information Retrieval oder bei der automatischen Anfrageerweiterung mittels Thesauri eingesetzt werden.

Schließlich wird in Kapitel 6 die Einbindung MorphoSaurus-System als Modul in das Informationssystem der Hautklinik Freiburg beschrieben. Dort soll die Alltagstauglichkeit des Systems unter Beweis gestellt und die Benutzerzufriedenheit für das System ermittelt werden. Außerdem werden einige Herausforderungen besprochen, die die Portierung eines wissenschaftlichen Systems in ein anwendungsbezogenes Umfeld mit sich bringt.

In Kapitel 7 werden abschließend Stärken und Schwächen des MorphoSaurus-Systems rekapituliert und ein Ausblick auf weitere Trends und Entwicklungen geben, die im Rahmen des MorphoSaurus-Projektes verfolgt werden.

Kapitel 2

Das MorphoSaurus-System

The true signigficant elements of language are ... either words, significant parts of words, or word groupings. [Sapir1921]

2.1 Einleitung

Das Wort gilt als die grundlegende Einheit der Sprache. Die Definition dessen, was ein Wort ist, stellt dabei nach wie vor ein Problem der linguistischen Theorie [Bauer1983] dar, und die verschiedenen Sichtweisen auf Wörter scheinen sich teils zu ergänzen, teils zu widersprechen. Morphologisch gesehen ist ein Wort eine möglichst kleine sprachliche Einheit, die eine Bedeutung trägt und frei vorkommen kann. Aus syntaktischer Sicht sind Wörter Einheiten, die sich innerhalb eines Satzes verschieben, sich durch andere austauschen und sich durch das Einfügen weiterer Wörter voneinander trennen lassen. Semantisch gesehen zeichnen sich Wörter dadurch aus, dass sie im Allgemeinen eine Bedeutung tragen. In den meisten sprachverarbeitenden Computer-Systemen wird ein Wort in der Regel als eine durch Leerzeichen abgrenzbare Einheit einer Zeichenkette angesehen.

Semantisch gesehen erscheint die wortzentrierte Sicht auf die Sprache nicht immer sinnvoll. Betrachtet man den englischen Ausdruck *"inflammantion of the appendix"*, so wird seine Bedeutung durch die zwei lexikalischen, also inhaltliche Bedeutung tragenden Wörter *"inflammation"* und *"appendix"* repräsentiert. Allerdings lässt sich der Ausdruck sowohl im Englischen mit *"appendicitis"* als auch in anderen Sprachen durch ein einzelnes Wort repräsentieren, wie beispielsweise im Deutschen durch *"Blinddarmentzündung"*, im Holländischen durch *"blindedarmontsteking"* oder im Schwedischen *"blindtarmsinflammation"*. Es erscheint daher aus semantischer Sicht sinnvoll, nicht Wörter voneinander abzugrenzen, sondern Einheiten semantischer Atomarität. Dies können Wörter, signifikante Wortteile oder auch

Wortgruppen sein [Sapir1921]. So können in dem genannten Beispiel die Wortbestandteile *"appendix", "appendic", "blinddarm", "blindtarms", "blindedarm"* sowie *"inflammation", "itis", "ontsteking"* voneinander abgegrenzt werden. Ein Beispiel für einen semantisch atomaren Mehrwortausdruck ist *"Multiple Sklerose"*. Hier lässt sich der Sinn des Mehrwortausdruckes nicht ohne weiteres aus den einzelnen Wörtern herleiten, da es sich bei dem Begriff *"Multiple Sklerose"* um ein komplexes neurologisches Krankheitsbild mit einer Vielzahl unterschiedlicher Symptome handelt. Die wörtliche Bedeutung *"vielfache Gewebsverhärtungen"* beschreibt jedoch lediglich einen morphologischen Befund, der mit dem Krankheitsbild einhergeht.

Das MorphoSaurus-System[1] ist ein IR-System, welches ursprünglich für die multilinguale Dokumentenrecherche (engl. *Cross-Language Information Retrieval, CLIR*) im biomedizinischen Kontext entwickelt wurde, bei der Benutzeranfragen und Dokumentenkollektionen in unterschiedlichen Sprachen verfasst sind [Markó et al.2005a, Markó2007]. Dieses System trägt der semantisch orientierten Sichtweise auf die Sprache Rechnung und führt sogenannte *Subwörter* als semantisch atomare Einheiten ein. Diese sind in sprachspezifischen Lexika definiert und werden über einen Thesaurus sprachübergreifend miteinander verknüpft. Basierend auf dem zugrunde liegenden Modell von MorphoSaurus lassen sich eine Reihe von Applikationen entwickeln, von denen einige im Laufe dieser Arbeit vorgestellt werden. Dieses Kapitel gibt einen Überblick über die wesentlichen Konzepte des MorphoSaurus-Systems und beschreibt die Realisierung dieser Konzepte in der biomedizinischen Domäne.

2.2 MorphoSaurus-Subwörter

Subwörter sind definiert als atomare konzeptuelle oder linguistische Einheiten, deren Bedeutung nicht von kleineren bedeutungstragenden Einheiten abgeleitet werden kann ([Markó et al.2005a]). Subwörter ähneln somit Morphemen, die als Klassen kleinster bedeutungstragender Einheiten ([Helbig1998]) definiert sind. Im Gegensatz zu Morphemen werden bei Subwörtern jedoch auch Morphem-Kombinationen als atomare Einheiten, also als einzelne Subwörter definiert, wenn deren Bedeutung sich nicht aus den einzelnen Bestandteilen ableiten lässt. Beispielsweise lassen sich die Begriffe *"Bauchspeicheldrüse"* oder *"Schenkelhals"* nicht unmittelbar aus den Wortbestandteilen *"Bauch"*, *"Speichel"* und *"Drüse"* bzw. *"Schenkel"* und *"Hals"* herleiten und werden somit als eigene Subwörter (*"bauchspeicheldrüs"* bzw. *"schen-*

[1] Der Name MorphoSaurus entstand als Akronym für *MORPHem theSAURUS*.

kelhals")[2] definiert. Damit bleiben in Subwörtern Wortbedeutungen bestehen, die durch das Trennen in mehrere Morpheme verloren gehen.

Das MorphoSaurus-Modell erlaubt die Definition von *Mehrwort-Subwörtern*[3]. Dies ist insbesondere der Fall, wenn Mehrwortbegriffe einen festen Begriff charakterisieren, der bei der Trennung der Wörter des Mehrwortbegriffes verloren ginge. Beispiele für Mehrwort-Subwörter sind *"vitamin c"* oder das englische *"yellow fever"* (*"Gelbfieber"*).

Subwörter werden durch drei Attribute charakterisiert, durch ihre *Sprachzugehörigkeit*, ihre *Domänenzugehörigkeit* und durch ihre Zugehörigkeit zu einem bestimmten *Subworttyp*. Diese Attribute werden im Folgenden beschrieben.

2.2.1 Sprachzugehörigkeit

Subwörter sind sprachspezifisch. Für das Deutsche werden beispielsweise die Subwörter *"herz"* und *"nier"* definiert, für ihre englischen Äquivalente die Subwörter *"heart"* und *"kidney"*. Subwörter aus Sprachen, die im gewöhnlichen Sprachgebrauch einer anderen Sprache vorkommen, werden in beiden Sprachen als Subwörter definiert. So werden die englischen Wörter *"pouch"* (*"Speichermöglichkeit für Stuhl"*) oder *"feedback"* (*"Rückmeldung"*) auch im Deutschen als Subwörter definiert. Das MorphoSaurus-Modell definiert $\mathcal{L} := \{Englisch(EN), Deutsch(DE), Portugiesisch(PT), \ldots\}$ als die Menge der betrachteten Sprachen.

2.2.2 Domänenzugehörigkeit

Subwörter werden immer im Kontext einer bestimmten Domäne definiert. Beispielsweise hat der Begriff *"Leiste"* im Handwerk (*längliches Bauteil*) eine andere Bedeutung als in der Medizin (*anatomische Lokalisation*). Daher muss zu jedem Subwort immer auch die Domäne angegeben werden, in der dieses Subwort definiert ist. Die Menge aller betrachteten Domänen im MorphoSaurus-Modell ist definiert als $\mathcal{D} = \{KlinischeMedizin(CM), Rechtswesen(J), \ldots\}$.

[2]Subwörter werden per conventionem in Kleinbuchstaben geschrieben.

[3]Der Begriff Subwort ist in diesem Zusammenhang irreführend, da sich Subwörter in diesen Fällen über die Wortgrenzen hinaus erstrecken können, also längere Einheiten als einzelne Wörter darstellen.

2.2.3 Typzugehörigkeit

Ähnlich wie bei Morphemen können bei Subwörtern verschiedene Subworttypen unterschieden werden. Subwörter, die selbständig als Wort vorkommen können, werden als *frei* bezeichnet. Ein Beispiel für ein freies Subwort ist *"herz"*. Subwörter, die nicht frei vorkommen, wie *"-s"*, *"-st"* oder *"-itis"* werden *gebunden* genannt. Ein Subwort, das eine Grundbedeutung ausdrückt, heißt *Stamm*. Nach ihrer Platzierung relativ zum *Stamm* unterscheidet man weiterhin das (echte) *Präfix*, das (echte) *Suffix*, das *Infix*[4] und die *Invariante*.

- **Stämme** (ST) wie *"leber"*, *"schilddrues"*, "enferm" (spanisch für *"krank"*) und *"append"*, drücken die Grundbedeutung eines Wortes aus. Stämme können miteinander verknüpft werden (mit oder ohne Affixe) sowie durch Präfixe und Suffixe ergänzt werden (wie in *"Blinddarm⊕entzünd⊕ung"*)[5]. Stämme können auch ohne Affixe auftreten (wie int *"herz"* oder *"muskel"*).

- **Präfixe** (PF) wie *"hyper-"*, *"peri-"*, *"contra-"*, *"circun-"*, *"hemi"* und *"an"* stehen vor einem Stamm (wie in *"Hyper⊕tonie"*) oder einem Präfix (wie in *"Hemi⊕an⊕opsie"*).

- **Echte Präfixe** (Proper Prefixes, PPF) wie *"omo-"* (*"Schulter"*), *"down-"* und *"hemi-"* sind Präfixe, vor die kein anderes Subwort gesetzt werden kann. Sie wurden eingeführt, um die Zahl der Fehlzerlegungen, die durch kurze Präfixe entstehen, zu minimieren (wie in *"Hyp⊕omo⊕bilität"*). Durch die Einführung von *"omo-"* als echtem Präfix ist eben genannte Fehlzerlegung nicht mehr möglich und die korrekte Zerlegung *"Hypo⊕mobil⊕ität"* wird gefunden.

- **Infixe** (IF) wie *"-o-"* in *"gastr⊕o⊕intestinal"* oder *"r"* in *"Hernio⊕r⊕raphie"* werden aus phonetischen Gründen zwischen zwei Stämme eingefügt, um diese miteinander zu verknüpfen.

- **Suffixe** (SF) können einem Stamm oder einem anderen Suffix folgen. Dazu gehören sowohl bedeutungstragende (lexikalische) Suffixe wie *"-itis"* in *"pankreat⊕itis"* oder *"-logy"* in *"cardio⊕logy* als auch nicht bedeutungstragende (grammatische) Suffixe (wie *"-ion"* in *"Reakt⊕ion"*).

- **Echte Suffixe** (Proper Suffixes, PSF), wie *"-ing"* in *"absorb⊕ing"* (engl. *"absorbierend"*), *"-alities"* in *"common⊕alities"* (engl. *"Gemeinsamkeiten"*) ste-

[4](Echte) Suffixe und Präfixe sowie Infixe werden zusammen als Subwort-*Affixe* bezeichnet.
[5]Das Symbol ⊕ bezeichnet den Verknüpfungsoperator.

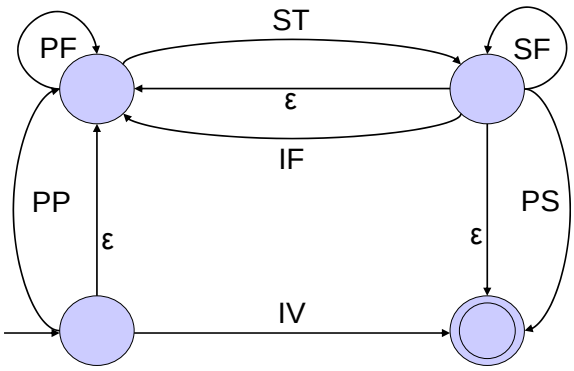

Abbildung 2.1: Darstellung der erlaubten Folgen von Subworttypen als endlicher Automat.

hen immer am Ende eines Wortes und können somit nicht vor ein anderes Subwort gestellt werden.

- **Invarianten** (IV) stellen Subwörter dar, die nicht den Wortbildungsphänomenen wie Flexion, Derivation und Komposition unterworfen werden. Hierzu gehören zum einen kurze Wörter wie *"Gen", "Ei"* und *"Ion"*, die als Stämme zu vielen Fehlerzerlegungen führen (wie in *"Abdukt⊕ion"*), aber auch einige Eigennamen (*"Fleming"*) oder Abkürzungen wie *"WHO"* (*"World Health Organization"*) oder *"AIDS"* (*"Acquired Immuno-Defiency Syndrome"*).

Im MorphoSaurus-Modell sind die aufgeführten Subworttypen als Menge $\mathcal{T}$:={ST,PF,PPF,IF,SF,PSF,IV} definiert. Die Regeln, die die möglichen Kombinationen von Subworttypen beschreiben, sind in Abbildung 2.1 in Form eines endlichen Automaten dargestellt. Ein Wort startet optional mit einem (echten) Präfix, gefolgt von mindestens einem Stamm. Auf Stämme können weitere Stämme folgen, die durch optionale Präfixe, Infixe oder Suffixe verknüpft sind. Als Letztes folgen die ebenfalls optionalen echten Suffixe.

Mit der Einführung von Sprache $\mathcal{L}$, Domäne $\mathcal{D}$ und Typ $\mathcal{T}$ als charakteristische Attribute von Subwörtern erfolgt nun die formale Definition eines Subwortes S als Element des folgenden Quadrupels:

$$\mathcal{S} \subset \mathcal{SID} \times \mathcal{T} \times \mathcal{L} \times \mathcal{D},$$

mit $\mathcal{SID} := \{gastr, hepat, enferm, de, anti, itis, \ldots\}$ als nichteindeutiger sprechender Bezeichner des jeweiligen Subwortes. Typische Beispiele für Subwörter nach dieser Notation sind *(inflamm,ST,DE,CM)*, *(herz,ST,DE,CM)* oder *(logy,SF,EN,CM)*. Aus Gründen der Übersichtlichkeit wird im Folgenden jedoch auf

diese formale Notation verzichtet und lediglich der Subwort-Bezeichner $\mathcal{SID}$ (*"inflamm", "herz", "logy"*) angegeben, wenn von Subwörtern die Rede ist.

Subwörter sind in sprachspezifischen Subwort-Lexika $\mathcal{LE}$ organisiert. Tabelle 2.1 auf Seite 29 zeigt im oberen Teil exemplarisch ein solches Subwort-Lexikon für die deutsche und englische Sprache (oben-rechts bzw. oben-links) in der Domäne der *Klinischen Medizin (CM)*.

2.3 MorphoSaurus-Äquivalenzklassen

Subwörter können auf verschiedene Arten miteinander in Beziehung stehen. So können die deutschen Subwörter *"inflamm"* und *"itis"* sowie das englische Subwort *"itic"* als Synonyme angesehen werden. Synonyme Subwörter werden in sprachunabhängige Gruppen zusammengeführt, die als Äquivalenzklassen bezeichnet werden und durch so genannte eindeutige MIDs (Morphosaurus Identifiers) repräsentiert werden. Definitionsgemäß beginnen MIDs immer mit dem Symbol '#', gefolgt von einem englischen Subwort aus dieser Äquivalenzklasse[6]. Die oben genannten Subwörter werden in der Äquivalenzklasse mit der MID "#inflamm" zusammengefasst. Anschaulich gesprochen stellen MIDs Wörter einer Interlingua dar, durch die verschiedene Sprachen in eine gemeinsame sprachunabhängige Repräsentation überführt werden können. Beispielsweise werden der deutsche Ausdruck *"Akute Blinddarmentzündung"* und der englische Ausdruck *"acute inflammation of the appendix"* durch dieselbe Interlingua repräsentiert: *"#acut #appendix #inflamm"*[7].

Im Folgenden wird die Beziehung zwischen Subwörtern und Äquivalenzklassen formal definiert. $\mathcal{M}$ sei dazu definiert als die Menge aller Äquivalenzklassen. Der Übergang von Subwörtern zu Äquivalenzklassen wird als semantische Normalisierung (SN) bezeichnet. Dieser Übergang kann formal aufgefasst werden als Relation $\mathcal{F}_{\mathcal{SN}}$, in der für jedes Subwort (sid, t, l, d) $\in \mathcal{S}$ genau eine Äquivalenzklasse m $\in \mathcal{M}$ existiert, so dass ((id, t, l, d), m) $\in \mathcal{F}_{\mathcal{SN}}$ gilt. Dies bedeutet, dass für jedes einzelne oder für mehrere synonyme Subwörter genau eine Äquivalenzklasse existiert, auf die die Subwörter abgebildet werden.

Mit Hilfe der Funktion $\mathcal{F}_{\mathcal{SN}}$ können nun Sublexika definiert werden, die für die formale Definition von Äquivalenzklassen eine Rolle spielen. Das Sublexikon $\mathcal{LE}_{\#abdomen}$ enthält alle Subwörter, die als Funktionswert $\mathcal{F}_{SN}$ die Äquivalenzklasse mit der MID "*#abdomen*" besitzen. Anders ausgedrückt, sind in dem Sublexikon

[6] Ist kein englisches Subwort in der Äquivalenzklasse enthalten, wird nach einer bestimmten Reihenfolge ein Subwort aus einer anderen Sprache gewählt.

[7] Die MID-Reihenfolge wird hier nicht berücksichtigt.

$\mathcal{LE}_{\#abdomen}$ sowohl das Subwort *"abdomen"* als auch alle seine Synonyme wie *"bauch"* enthalten. Formal ergibt sich folgende logische Äquivalenz:

$$\{(abdomen, ST, E, CM), (belly, ST, E, CM), (bauch, ST, E, CM)\} \in \mathcal{LE}_{\#abdomen} \Leftrightarrow$$
$$\#abdomen = \mathcal{F}_{SN}(abdomen, ST, EN, CM) = \mathcal{F}_{SN}(belly, ST, EN, CM) = \mathcal{F}_{SN}(bauch, ST, DE, CM)$$

Äquivalenzklassen können dahingehend unterteilt werden, ob sie im Kontext der Dokumentenrecherche eine relevante semantische Bedeutung tragen (*lexikalische Äquivalenzklassen*) oder nicht (*grammatische Äquivalenzklassen*). Äquivalenzklassen, in denen Stämme enthalten sind, werden im Allgemeinen als bedeutungstragend angesehen. Infixe sowie zahlreiche Invarianten, Präund Suffixe hingegen haben lediglich eine grammatische Funktion und lassen die semantische Bedeutung in Wörtern weitgehend unbeeinflusst. Dazu zählen Hilfsverben (wie *"bin", "war", "gewesen"*), Suffixe (wie *"-ion", "-s", "-iert"*) oder klassische Stoppwörter (wie *"nun", "weil", "daher"*). Alle grammatischen Äquivalenzklassen werden durch ein leeres MID (ε) ausgedrückt und werden in Anwendungen nicht weiter berücksichtigt. Die Entscheidung, ob eine Äquivalenzklasse bedeutungstragend ist oder nicht, ist nicht immer leicht zu treffen: Während das Suffix *"-itis"* im Allgemeinen als bedeutungstragend angesehen wird, fällt die Entscheidung bei den Suffixen *"-ulus"* (in *"nod⊕ulus"*) und *"-fähig"* (in *"reaktions⊕fähig"*) bereits schwerer.

Während der Implementierung des MorphoSaurus-Modells stellte sich heraus, dass sich zwei Relationen zwischen Äquivalenzklassen als besonders sinnvoll erwiesen, um bestimmte Zusammenhänge zwischen den Wörtern der Interlingua besser ausdrücken zu können. Entsprechend wurden diese Relationen in das Modell aufgenommen und werden an dieser Stelle formal eingeführt. Die zwei Relationen sind die *expandsTo*-Relation und die *hasSense*-Relation:

- $expandsTo \subset \mathcal{M} \times \mathcal{M}$: Die Menge $S_1 := \{(m_0, m_1), (m_0, m_2), \ldots, (m_0, m_n)\} \in expandsTo$ (mit $m_{0,\ldots,n} \in \mathcal{M}$ und $|S_1| \geq 2$) setzt eine Äquivalenzklasse m_0 zu mindestens zwei anderen Äquivalenzklassen in Beziehung. Sie drückt aus, dass die Äquivalenzklasse m_0 nicht eine inhaltlich atomare Bedeutung trägt, wie das im Allgemeinen von Äquivalenzklassen gefordert ist, sondern dass m_0 aus zwei oder mehreren bedeutungstragenden Einheiten besteht. Welche bedeutungstragenden Einheiten dies sind, wird mit der *expandsTo*-Relation definiert. Diese Relation wird aus folgenden Gründen eingeführt:

  1. Sehr kurze Morpheme können nicht als Subwörter zugelassen werden, da die Gefahr von Fehlzerlegungen sehr groß ist (bspw. die Zerlegung

von *"eigen"* in *"ei⊕gen"*). Als Konsequenz müssen Wörter, die diese kurzen Morpheme als Wortbestandteile enthalten (wie *"Ei⊕sprung"*, *"Ei⊕bläschen"*), als eigene Subwörter aufgenommen werden und deren semantische Dekomposition durch eine $expandsTo$-Relation kodiert werden. Das folgende Beispiel demonstriert die $expandsTo$-Relation anhand des Subwortes *"myalg"*. *"myalg"* wird als eigenes Subwort definiert, da sowohl *"my"* als auch *"alg"* zu viele Fehlzerlegungen verursachen. Formal betrachtet wird die $expandsTo$-Relation nun folgendermaßen ausgedrückt. Seien folgende Subwort-Lexika für *"myalg"*, *"muscle"* und *"pain"* definiert:

$$\begin{aligned}
&\{(myalg, ST, EN, CM), (mialg, ST, SP, CM)\} && \in \mathcal{LE}_{\#myalg}\\
&\{(muscle, ST, EN, CM), (muscul, ST, SP, CM)\} && \in \mathcal{LE}_{\#muscle}\\
&\{(pain, ST, EN, CM), (algia, SF, SP, CM)\} && \in \mathcal{LE}_{\#pain}
\end{aligned}$$

Mit der folgenden $expandsTo$-Relation kann nun auf die Definition von "my" und "alg" als Subwörter verzichtet werden:

$$(\#myalg, \#muscle), (\#myalg, \#pain) \in expandsTo$$

2. Ein zweiter Grund für die $expandsTo$-Relation ist gegeben, wenn in einer Sprache ein nicht zerlegbares, atomares Subwort eine zusammengesetzte Bedeutung in einer anderen Sprache besitzt. Zum Beispiel:
$$(\#esparadrap[8], \#adhesiv), (\#esparadrap, \#tape) \in expandsTo$$
3. In Wörtern mit mehreren Morphemen werden einzelne Buchstaben ausgelassen (sogenannte Ellipsen), wie in *"urinalyis"* (*"Urinanalyse"*):
$$(\#urinalys, \#urin), (\#urinalys, \#analys) \in expandsTo$$

- $hasSense \subset \mathcal{M} \times \mathcal{M}$: Die Menge $S_2 := \{(m_0, m_1), (m_0, m_2), \ldots, (m_0, m_n)\} \in hasSense$ (with $m_{0,\ldots,n} \in \mathcal{M}$ und $|S_2| \geq 2$) verknüpft eine mehrdeutige Äquivalenzklasse mit mindestens zwei weiteren Äquivalenzklassen, die die verschiedenen Bedeutungen dieser mehrdeutigen Äquivalenzklasse repräsentieren. Das spanische Subwort *"lobo"* beispielsweise ist doppeldeutig und bezeichnet zum einen den *"Wolf"*, zum anderen den anatomischen Begriff *"Lappen"*. Die Äquivalenzklassen mit den Sublexika

$$\begin{aligned}
&\{(lobo, IV, SP, d), (lobos, IV, SP, d)\} && \in \mathcal{LE}_{\#lobo}\\
&\{(wolf, ST, EN, d), (wolves, ST, EN, d)\} && \in \mathcal{LE}_{\#wolf}\\
&(lob, ST, EN, d) && \in \mathcal{LE}_{\#lobe}
\end{aligned}$$

[8] Portugiesisch für *"Pflaster"*

werden daher über die *hasSense*-Relation miteinander in Beziehung gesetzt. In der vereinfachten Darstellung werden diese Äquivalenzklassen durch die jeweiligen MIDs angegeben, die durch ein Semikolon getrennt sind:

$$(\#\{wolf; lobe\}, \#wolf), (\#\{wolf; lobe\}, \#lobe)) \in hasSense$$

Abgesehen von *expandsTo*/*hasSense*-Relationen werden keine weiteren semantischen Relationen eingeführt. Insbesondere wird auf hierarchische Relationen wie Hyperonymie und Hyponymie (Ober- und Unterbegrifflichkeit) oder Meronymie (Teil-Ganzes-Beziehungen) verzichtet. Diese sind bereits in anderen Thesauri wie UMLS([Nelson et al.2002]), WordNet[Fellbaum1998], MeSH[9] oder [Côté et al.1993] definiert. Daher ist ein vordringliches Ziel, keine zu diesen Thesauri parallele Hierarchie-Struktur aufzubauen, sondern diese Ressourcen zu nutzen und auf die Einträge dieser Thesauri abzubilden [Daumke et al.2003, Markó et al.2003].

Die in diesem Kapitel bisher eingeführten Definitionen beinhalten MIDs als Bezeichner von Äquivalenzklassen, $\mathcal{LE}_{MID}$ als Sublexika dieser Äquivalenzklassen sowie die *expandsTo*/*hasSense*-Relationen zur semantischen Verknüpfung zwischen Äquivalenzklassen. Mit Hilfe dieser Definitionen lassen sich die Äquivalenzklassen $\mathcal{M}$ nun formal als Elemente des folgenden Quadrupels definieren:

$$\mathcal{M} \subset \mathcal{MID} \times \mathcal{LE}_{MID} \times expandsTo \times senseOf$$

Ein Beispiel für eine typische Äquivalenzklasse ist folgendes Quadrupel:

$$(\#urinalys, ((urinalys, ST, EN, CM), (urinalis, ST, PT, CM)),$$
$$((\#urinalys, \#urin), (\#urinalys, \#analys)), (\varepsilon)),$$

mit

$\#urinalys$	$\in \mathcal{MID}$
$(urinalys, ST, EN, CM), (urinalis, ST, PT, CM)$	$\in \mathcal{LE}_{MID}$
$(\#urinalys, \#urin), (\#urinalys, \#analys)$	$\in expandsTo$
(ε)	$\in senseOf$

Äquivalenzklassen sind im sprachunabhängigen *Subwort-Thesaurus* $\mathcal{T}$ organisiert. Tabelle 2.1 (unten) stellt einen solchen Thesaurus exemplarisch dar. Darin wird der semantisch zusammengesetzte Begriff *"hyoid"* (*"Zungenbein"*) in die Begriffe *"Zunge"* (engl. *"tongue"*) und *"Knochen"* (engl. *"bone"*) zerlegt. Außerdem wird die Mehrdeutigkeit des Subwortes *"bruch"* in die Bedeutungen *"fraktur"* (*"Knochenbruch"*) und *"hernie"* (*"Weichteilbruch"*) aufgelöst.

[9] `http://www.nlm.nih.gov/mesh/`, eingesehen im Februar 2007

Tabelle 2.1: Beispiel eines minimalen deutschen und englischen Lexikons sowie des dazugehörigen Thesaurus.

Englisches Subwort-Lexikon	**Deutsches Subwort-Lexikon**
$\mathcal{LE}_{EN} := \{$	$\mathcal{LE}_{DE} := \{$
(a, IV, EN, CM),	$(zunge, ST, DE, CM)$,
$(hyoid, ST, EN, CM)$,	(n, SF, DE, CM),
$(fracture, ST, EN, CM)$,	$(bein, ST, DE, CM)$,
(is, IV, EN, CM),	$(bruech, ST, DE, CM)$,
$(rare, ST, EN, CM)$,	(e, ST, DE, CM),
$(phenomenon, ST, EN, CM)$,	$(sind, IV, DE, CM)$,
$(that, IV, EN, CM)$,	$(selten, ST, DE, CM)$,
(may, IV, EN, CM),	$(ereigniss, ST, DE, CM)$,
$(result, ST, EN, CM)$,	(mit, IV, DE, CM),
(in, IV, EN, CM),	$(teils, IV, DE, CM)$,
$(signific, ST, EN, CM)$,	$(erheblich, ST, DE, CM)$,
(ant, SF, EN, CM),	(en, SF, DE, CM),
$(complic, ST, EN, CM)$,	$(komplik, ST, DE, CM)$,
$(ations, PS, EN, CM)$	$(ationen, SF, DE, CM)$
$\}$	$\}$

Subwort-Thesaurus
$\mathcal{T} := (\mathcal{MID}, \mathcal{LE}_{MID}, expandsTo, hasSense)$, mit
$\{\#hyoid, ((hyoid, ST, EN, CM), (hyoid, ST, DE, CM)),$
$\quad((\#hyoid, \#tongue), (\#hyoid, \#bone)), \varepsilon\}$
$\{\#bruch, (bruech, ST, DE, CM)), \varepsilon, ((\#bruch, \#fract), (\#bruch, \#herni))\}$
$\{\#tongue, ((tongue, ST, EN, CM), (zung, ST, DE, CM)), \varepsilon, \varepsilon\}$
$\{\#bone, ((bone, ST, EN, CM), (knochen, ST, DE, CM)), \varepsilon, \varepsilon\}$
$\{\#fract, ((fract, ST, EN, CM), (frakt, ST, DE, CM)), \varepsilon, \varepsilon\}$
$\{\#hernia, ((herni, ST, EN, CM), (herni, ST, DE, CM)), \varepsilon, \varepsilon\}$
$\ldots$

Durch die Verwendung sprachspezifischer Subwörter und sprachunabhängiger Äquivalenzklassen können eine Reihe lexikalischer Relationen abgebildet werden, was für die ein- und mehrsprachige Dokumentenrecherche einen besonderen Nutzen darstellt. Einige dieser Relationen wie Synonymie, $expandsTo$ und $hasSense$-Relation wurden bereits beschrieben. Zur Übersicht werden zum Abschluss dieses Kapitels noch einmal alle im MorphoSaurus-Modell ausdrückbaren Relationen in einer Übersicht dargestellt:

- **Synonymie**: Die englischen Suffixe *"itic"* und *"itis"* haben (in derselben Domäne) dieselbe Bedeutung wie der englische Stamm *"inflammation"*.
  $$\begin{aligned}\#inflamm &= \mathcal{F}_{SN}(inflamm, ST, EN, CM)\\ &= \mathcal{F}_{SN}(itic, SF, EN, CM)\\ &= \mathcal{F}_{SN}(itis, SF, EN, CM)\end{aligned}$$

- **Übersetzung**: Der deutsche Stamm *"entzuend"* und das französische Suffix *"itis"* haben die gleiche Bedeutung wie der englische Stamm *"inflamm"*.
  $$\begin{aligned}\#inflamm &= \mathcal{F}_{SN}(inflamm, ST, EN, CM)\\ &= \mathcal{F}_{SN}(entzuend, ST, DE, CM)\\ &= \mathcal{F}_{SN}(itis, SF, FR, CM)\end{aligned}$$

- **Koinzidenz**: Das Subwort *"mister"* beschreibt im Englischen eine *"männliche Person"*, im Spanischen und Portugiesischen hingegen bedeutet dieser Wortstamm *"Geheimnis, Mysterium"*.
  $$\begin{aligned}\#mister &= \mathcal{F}_{SN}(mister, ST, EN, CM)\\ \#myster &= \mathcal{F}_{SN}(mister, ST, SP, CM)\\ &= \mathcal{F}_{SN}(mister, ST, PT, CM)\end{aligned}$$

- **Domänenspezifität**: Die Namen *"sildenafil"* und *"viagra"* können im Bereich der *Klinischen Medizin* (CM) als Synonyme angesehen werden, jedoch nicht in der Domäne der *Pharmazeutischen Industrie* (PI).
  $$\begin{aligned}\#sildenafil &= \mathcal{F}_{SN}(sildenafil, ST, EN, CM)\\ &= \mathcal{F}_{SN}(viagra, ST, EN, CM)\\ &= \mathcal{F}_{SN}(sildenafil, ST, EN, PI)\\ \#viagra &= \mathcal{F}_{SN}(viagra, ST, EN, PI)\end{aligned}$$

- **Mehrdeutigkeit**: Das englische Wort *"head"* kann sich auf eine anatomische Lokalisation beziehen oder auf den Leiter einer Abteilung. Für doppeldeutige Subwörter werden eigene Äquivalenzklassen definiert. Diese können über die $senseOf$-Relation mit den eindeutigen Äquivalenzklassen nachträglich verknüpft werden:

$$\#\{leader; caput\} = \mathcal{F}_{SN}(head, ST, DE, CM)$$
$$\#leader \quad = \mathcal{F}_{SN}(leader, ST, DE, CM)$$
$$\#caput \quad = \mathcal{F}_{SN}(caput, ST, DE, CM)$$

- **Komposition**: Das deutsche Wort *"Myalgie"* ist eine Zusammensetzung aus den zwei elementaren Einheiten *"Muskel"* und *"Schmerz"*. Dennoch wird *"myalg"* als eigenes Subwort definiert, da die beiden Subwörter *"my"* und *"alg"* zu vielen Fehlzerlegungen führen. Über die "expandsTo"-Relation ist es jedoch möglich, die beiden Bedeutungsbestandteile anzugeben:
  $\#myalg \; = \mathcal{F}_{SN}(myalg, ST, DE, CM)$
  $\#muscle = \mathcal{F}_{SN}(muskel, ST, DE, CM)$
  $\#pain \quad = \mathcal{F}_{SN}(schmerz, ST, DE, CM)$

2.4 Morpho-Semantische Indexierung

Die sprachspezifischen Subwort-Lexika sowie der Thesaurus stellen die lexikalischen Komponenten des MorphoSaurus-Systems dar. Zusätzlich besteht das System aus Algorithmen, mit denen Texte in Äquivalenzklassen überführt werden können. Diese Algorithmen bilden in ihrer Gesamtheit den Subwort-Indexierer und implementieren eine dreistufige Funktion, die als morpho-semantische Indexierung bezeichnet wird. Diese Funktion wird in den folgenden Abschnitten näher beschrieben. Abbildung 2.2 stellt die drei Stufen der morpho-semantischen Indexierung für einen englischen und einen deutschen Satz basierend auf den Lexika und dem Thesaurus aus Tabelle 2.1 grafisch dar.

2.4.1 Orthografische Normalisierung

Ein Präprozessor verwandelt den Eingabetext zunächst in Kleinbuchstaben und 7-bit ASCII-Code (Abbildung 2.2 (oben)). Dabei werden sprachspezifische Ersetzungsregeln angewendet, im Deutschen beispielsweise *'ß'* → *'ss'*, *'ä'* → *'ae'*, *'ö'* → *'oe'*, *'ü'* → *'ue'* und im Portugiesischen *'ç'* → *'c'*, *'ú'* → *'u'*, *'õ'* → *'o'*[10]. Zusätzliche Transformationsregeln sind durch Idiosynkrasien der medizinischen Terminologie motiviert, wie im Deutschen *'ca'* → *'ka'*, *'co'* → *'ko'*, *'cu'* → *'ku'*, *'ce'* → *'ze'*, *'ci'* → *'zi'*. Dadurch wird ein typisches Problem der Medizinterminologie gelöst [Brigl et al.1994],

[10]Eine Ausnahme wird für das Schwedische gemacht, bei dem die diakritischen Zeichen (å, ä und ö) erhalten bleiben, da sich die Semantik der Wörter bei Umwandlung in 7-bit ASCII-Code verändert (z.B. das Schwedische *vår* (*Frühling, unsere, unseres*) bzw. *var* (*Eiter, jeder, wo*)).

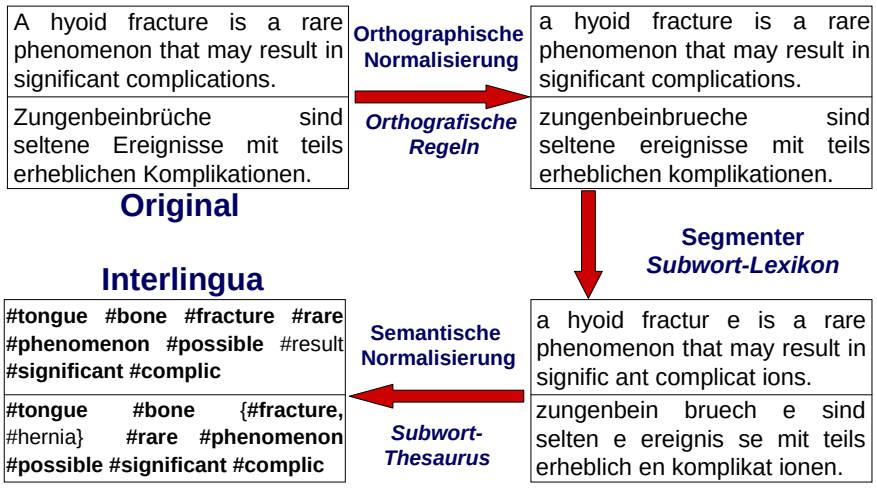

Abbildung 2.2: Die drei Stufen der morpho-semantischen Normalisierung in MorphoSaurus anhand eines deutsch-englischen Paralleltextes. Fett gedruckt sind die Übereinstimmungen auf Ebene der Äquivalenzklassen.

nämlich dass zahlreiche Wörter einerseits in Anlehnung an ihren lateinischen Ursprung mit 'c' geschrieben werden (in *"Carcinom"*), in anderen Fällen jedoch mit 'z' bzw. 'k' (in *"Carzinom"*, *"Karcinom"*, *"Karzinom"*).

2.4.2 Morphologische Segmentierung

Das MorphoSaurus-System segmentiert die orthografisch normalisierte Eingabe anschließend mit Hilfe des jeweiligen sprachspezifischen Subwort-Lexikons in eine Sequenz von Subwörtern (Abbildung 2.2 (rechts)). Die Segmentierung spielt sich dabei wie folgt ab: Jedes Eingabewort t der Länge n wird als Folge von Buchstaben $c_1, c_2, \ldots, c_n$ angesehen. Durch einen vorwärts und einen rückwärts laufenden Vergleichsprozess, der die Zeichenketten mit den Einträgen im entsprechenden Subwortlexikon vergleicht, werden die Eingabewörter in einzelne Subwörter zerlegt. Der vorwärts laufende Prozess startet mit der Zeichenkette $c_1, c_2, \ldots, c_k$ (mit $k{=}n$) und erniedrigt k schrittweise um eins, bis eine Teilzeichenfolge einen Treffer im Subwortlexikon ergibt. Die restliche Zeichenfolge $c_{k+1}, c_{k+2}, \ldots, c_n$ wird rekursiv auf die gleiche Art vorwärts und rückwärts durchlaufen, bis sämtliche Zeichen aus der Zeichenkette verarbeitet sind. Der rückwärts laufende Prozess startet ebenfalls mit der Zeichenkette $c_k, c_{k+1}, \ldots, c_n$ (mit $k = 1$) und erhöht k dann schrittweise um eins, bis das Teilwort mit einem Subwort aus dem Lexikon übereinstimmt. Das restliche Teilwort wird wiederum rekursiv vorwärts und rückwärts durchlaufen.

Mit Hilfe eines endlichen Automaten (siehe Abbildung 2.1) werden die möglichen Zerlegungen anschließend auf ihre morphologische Plausibilität überprüft, so darf

ein Wort beispielsweise nicht mit einem Suffix beginnen. Falls mehrere Zerlegungen gültig sind, werden verschiedene Heuristiken angewendet, um die richtige Zerlegung zu ermitteln. Falls keine gültige Segmentierung gefunden werden kann, werden alle Stämme mit vier oder mehr Buchstaben extrahiert und die restlichen Buchstaben verworfen. Wenn keine derartigen Stämme extrahiert werden können, wird das ursprüngliche Wort zurückgegeben[11].

2.4.3 Semantische Normalisierung

Jedes bei der morphologischen Segmentierung erzeugte Subwort wird im dritten Schritt, der *Semantischen Normalisierung*, durch sein entsprechendes MID ersetzt (Abbildung 2.2 (unten)), und der Text somit in eine sprachunabhängige Darstellung überführt. Sind für diese MIDs im Subwort-Thesaurus *expandsTo*-Relationen definiert, werden diese MIDs ersetzt durch diejenigen, auf welche die *expandsTo*-Relationen zeigen (beispielsweise wird *"#hyoid"* ersetzt durch *"#tongue #bone"*).

Disambiguierung mehrdeutiger Äquivalenzklassen

Im Falle von MIDs, für die *senseOf*-Relationen definiert sind (z.B. $\#\{wolf; lobe\}$), wird mittels eines Disambiguierers [Markó et al.2005b] versucht, das mehrdeutige MID durch die in diesem Kontext angesprochene Bedeutung zu ersetzen. Der Disambiguierer enthält dabei eine Liste von eindeutigen MID-Paaren, die aus Trainingsdaten ermittelt wurden und in denen angegeben ist, wie häufig diese Paare in den Trainingskorpora gemeinsam auftreten. Zwei MIDs gelten dabei als gemeinsam aufgetreten, wenn sie beide innerhalb einer Fenstergröße von vier MIDs erscheinen.

Werden nun zur Laufzeit mehrdeutige MIDs ermittelt, so bestimmt der Disambiguierer anhand der Maximum-Likelihood-Methode, welche Bedeutung in diesem Kontext die wahrscheinlichste ist. Als Kontext können wiederum die vier umgebenden MIDs betrachtet werden. Die Verwendung eines größeren Kontexts ist möglich, wirkt sich jedoch negativ auf die Laufzeit des Systems aus. Falls keine Disambiguierung möglich ist, wird das mehrdeutige MID ausgegeben.

Das Ergebnis der dreistufigen Prozedur ist ein morpho-semantisch normalisiertes Dokument in einer sprachunabhängigen Darstellung (Abbildung 2.2 (unten links)). Im Beispiel fett dargestellt sind alle MIDs, die sowohl im englischen als auch im

[11] Dieses Vorgehen hat sich besonders in monolingualen Retrieval-Szenarien als sinnvoll erwiesen. Auch bei multilingualen Szenarien bleiben durch dieses Vorgehen zum Beispiel Eigennamen erhalten und stehen den Applikationen weiter zur Verfügung.

deutschen Text auftreten. Deutlich wird die große Ähnlichkeit der beiden Texte auf Ebene der morpho-semantisch normalisierten Darstellung.

2.5 Implementierung des Subwort-Modells für die Biomedizin

Basierend auf dem in Abschnitt 2.2 beschriebenen MorphoSaurus-Modell wurde eine Implementierung dieses Modells für die Domäne $\mathcal{D}$ der *Klinischen Medizin (CM)* und die Sprachen $\mathcal{L} := \{EN, DE, PT, FR, SP, SE\}$ (Englisch, Deutsch, Portugiesisch, Französisch, Spanisch, Schwedisch) erstellt. Dieses Kapitel beschreibt die Strategien für die Erstellung und die Pflege der Subwort-Lexika und des Thesaurus. Als grafische Benutzeroberfläche wurde ein Web-basierter Editor entwickelt, der dem internationalen Team von Lexikographen die parallele Bearbeitung der lexikalischen Ressourcen an unterschiedlichen Orten ermöglicht (siehe Abschnitt 2.5.3). Wie die folgenden Abschnitte verdeutlichen, ist nicht immer unmittelbar eindeutig, welche Zeichenfolgen als Subwörter definiert werden sollen, und bisweilen führen verschiedene Formen der Implementierung zu sinnvollen Zerlegungen in Äquivalenzklassen. Um eine möglichst hohe Übereinstimmung zwischen Lexikographen zu erzielen, wurden strenge Leitlinien für die Implementierung erstellt.

2.5.1 Erstellung der Subwort-Lexika

Aus dem Web erhältliche umfangreiche Listen allgemeiner und domänenspezifischer Affixe dienen als Ausgangspunkt für die Erstellung der Subwort-Lexika. Um die Zahl kurzer Affixe, die bei der morpho-semantischen Normalisierung häufig zum Anstieg von Fehlzerlegungen führen, zu minimieren, werden gängige Affixkombinationen hinzugefügt (wie *"-igkeiten"*, *"-ectomies"*, *"-ationally"*). Anschließend wird mit der Ergänzung domänenspezifischer Stämme begonnen, die aus häufigkeitssortierten Wortlisten extrahiert werden. Dabei gibt das nach Entfernung der Affixe entstandene Teilwort häufig bereits einen ersten Hinweis auf einen gültigen Subwort-Stamm (wie in *"Entzünd⊕ung"*), wobei jedoch zahlreiche Ausnahmen bestehen (wie bei den inkorrekten Zerlegungen von *"Ent⊕fern⊕ung"*, *"Mag⊕en"*). Die automatische Präprozessierung der Wortlisten durch Verfahren wie regelbasierte Stammformbildung führen nicht zu einer wesentlichen Vereinfachung der Arbeit der Lexikographen, so dass die Extraktion von Subwort-Stämmen nach wie vor eine überwiegend intellektuelle Expertenarbeit darstellt.

Eine kontinuierliche Qualitätskontrolle wird durch die Erstellung neuer Wortlisten ermöglicht, die der morpho-semantischen Normalisierung unterzogen werden. Einige von ihnen dienen als Maß für die Güte der morpho-semantischen Normalisierung, indem die Rate korrekt segmentierter Wörter ermittelt wird, andere werden durch Lexikographen bearbeitet, um fehlende Stämme im Subwort-Lexikon zu ergänzen. Wann immer das Verknüpfen zweier Morpheme zu einem Bedeutungswandel führt, der über die kombinatorische Bedeutung dieser beiden Morpheme hinausgeht, wird diese Verknüpfung als eigenständiges Subwort definiert. Beispielsweise bedeutet der Begriff *"Neurose"* als eine spezielle Gruppe psychischer Störungen in der heutigen Medizin weit mehr als eine *"Nervenkrankheit"* (*"Neur⊕ose"*). Daher wird *"neuros"* als eigenständiges Subwort definiert. Gerade für Wörter lateinischen und griechischen Ursprungs existieren darüber hinaus eine Reihe so genannter Stammallomorphe (wie *"Corpus, Corpor⊕is"* oder *"Abdomen, Abdomin⊕al"*), die als zusätzliche Subwörter in das Lexikon übernommen werden. Dies verringert die Zahl kurzer Subwörter (wie *"korp"*), die zu Fehlzerlegungen bei der morpho-semantischen Indexierung führen können (wie in *"S⊕korp⊕ion⊕gift"*).

Nichteindeutige morphologische Segmentierung

Der in Kapitel 2.4.2 vorgestellte Zerlegungsalgorithmus zur morphologischen Segmentierung führt durch den parallelen Vorwärts- und Rückwärtsabgleich von Eingabewörtern mit dem Subwortlexikon gelegentlich zu mehreren möglichen Lesarten. Ein typisches Beispiel dafür ist das Wort *"Handlung"*, welches in die Subwörter *"hand⊕lung"* $\mapsto$ *"#hand #lung"* (1. Lesart) und *"handl⊕ung"* $\mapsto$ *"#handl"* (2. Lesart) zerlegt werden kann. Empirische Untersuchungen ergaben, dass im überwiegenden Teil der Fälle die Lesart mit dem längsten Subwort von links (sog. *left longest match*) die richtige Segmentierung darstellt, in diesem Beispiel also die zweite Lesart, da *"handl"* länger ist als *"hand"*. In einigen Fällen führt diese Regel jedoch auch zu Fehlsegmentierungen, wie Tabelle 2.2 verdeutlicht.

In diesen Fällen wird das Subwort-Lexikon nach pragmatischen Gesichtspunkten modifiziert. Im Folgenden werden die zur Verfügung stehenden Möglichkeiten für das englische Wort *"nephrotomy"* diskutiert.

- Die erste Möglichkeit besteht darin, das Subwort *"my"* aus dem Lexikon zu entfernen. Konsequenterweise müssen dann alle Komposita, die dieses Subwort enthalten, als Subwörter in das Lexikon aufgenommen werden (wie *"myalg"* oder *"myasthen"*) und über die $expandsTo$-Relation mit den einzelnen Bedeutungsbestandteilen verknüpft werden. Das setzt voraus, dass die Zahl der

Tabelle 2.2: Fehlsegmentierungen beim *left longest match*–Verfahren. In Klammern steht die korrekte Segmentierung.

Eingabewort	Morphologische Segmentierung	morpho-semantische Normalisierung
Gelenks-endoprothese	gelenk⊕send⊕o⊕prothes⊕e (gelenk⊕s⊕endo⊕prothes⊕e)	#joint #send #prothes (#joint #inner #prothes)
nephrotomy	nephr⊕oto⊕my (nephr⊕o⊕tomy)	#kidney #ear #muscle (#kidney #cut)
Schilddrüsen-geschwulst	schilddrues⊕enge⊕schwul⊕st (schilddrues⊕en⊕geschwulst)	#thyroid #close #gay (#thyroid #tumor)
Staatsexamen	staat⊕sex⊕amen (staat⊕s⊕examen)	#state #sex #amen (#state #exam)

Komposita überschaubar ist[12].

- Alternativ oder zusätzlich kann auch das Subwort *"oto"* aus dem Lexikon entfernt werden und durch Komposita wie *"otoskop"* oder *"otolarynx"* ersetzt werden.

- Als dritte Möglichkeit können zusätzliche Varianten des Subwortes *"nephr"* ergänzt werden, in diesem Falle das Subwort *"nephro"*.

- Schließlich kann auch eine Variante des Subwortes *"-tomy"* ergänzt werden (in diesem Fall *"-otomy"*), um die korrekte Zerlegung *"nephr⊕otomy"* zu erzeugen.

Welche der vier Möglichkeiten im Einzelfall verwendet wird, um die Fehlsegmentierung zu verhindern, lässt sich nicht allgemein beantworten. Die Entscheidung wird anhand verschiedener Einzelkriterien sowie nach umfangreichen Tests auf Testwortlisten getroffen. In diesem Fall erwies sich die vierte Möglichkeit als die beste.

Behandlung kurzer Subwörter

Je kürzer die Subwörter definiert werden, desto größer ist die Wahrscheinlichkeit, dass diese Buchstabenkombination zufällig in Wörtern auftaucht (wie *"ei"* und *"gen"* in *"eigen"*). Daher werden Regeln benötigt, nach denen entschieden werden kann, ob ein Subwort in das Lexikon aufgenommen wird oder nicht. Eine gängige Methode

[12]Die Zahl der Komposita wird anhand großer medizinischer Wortlisten ermittelt, die die medizinische Domäne umfangreich abdecken.

stellt die Untersuchung von Wortlisten dar, deren Wörter diese Buchstabenkombination enthalten. Zwei Szenarien sind möglich:

- **Die Anzahl falscher Treffer ist hoch:** Ein typisches Beispiel ist die Zeichenfolge *"ei"*. In den allermeisten Wörtern tritt diese Zeichenfolge nicht als Wortbestandteil mit der Bedeutung *"Ovum"* auf. In diesen Fällen werden die Subwörter *"ei"* und *"eier"* als Invarianten in das Subwortlexikon eingefügt, so dass die Eingabewörter *"Ei"* und *"Eier"* in Texten richtig erkannt werden. Außerdem werden alle Komposita, die als Wortbestandteil die Sequenz *"ei"* im Sinne von *"Ovum"* besitzen, als Subwörter im Lexikon definiert (wie *"eisprung"* oder *"eizell"*).

- **Die Anzahl korrekter Treffer ist hoch:** Wenn der umgekehrte Fall eintritt, nämlich dass eine Buchstabenfolge in den meisten Wörtern ein sinnvolles Subwort repräsentiert und nicht als zufälliger Treffer erscheint, wird eine andere Strategie verfolgt. Das entsprechende Subwort wird in diesen Fällen als Stamm in das Lexikon aufgenommen. Anschließend werden Wortlisten, die die entsprechende Buchstabenkombination enthalten, der morpho-semantischen Normalisierung unterzogen. In allen Fällen, in denen es durch die kurzen Subwörter zu Fehlzerlegungen kommt, wird nach einer praktischen Lösung gesucht und zusätzliche Subwort-Varianten eingefügt. Als Beispiel dient das Subwort *"oto"*, das als Subwort aufgenommen wird, da es in zahlreichen Wörtern als Wortbestandteil auftritt (wie in *"Otosklerose"*, *"ototoxisch"*, *"Otolayngoskopie"*). Treten bei Tests auf Wortlisten Fehlzerlegungen wie *"nephr⊕oto⊕my"* auf, werden Subwort-Varianten ergänzt, wie in diesem Falle *"-otomy"*, und die korrekte Zerlegung *"nephr⊕o⊕tomy"* wird ermöglicht.

2.5.2 Erstellung des Subwort-Thesaurus

Bei der Ergänzung eines neuen Subwortes in die Subwort-Lexika wird gleichzeitig eine neue Äquivalenzklasse angelegt, in deren Sub-Lexikon als einziger Eintrag das eben eingegebene Subwort enthalten ist. Frühzeitig wird nun damit begonnen, synonyme Subwörter in gemeinsamen Äquivalenzklassen zusammenzufassen. Dies geschieht mit der medizinischen Expertise der Lexikographen unter Zuhilfenahme verschiedener mono- und multilingualer lexikalischer Ressourcen wie UMLS [Lindberg et al.1993, Humphreys et al.1998], Leo (Link Everything Online) Dictionary[13] oder WordNet [Fellbaum1998]. Lexikographen geben Subwörter in der Re-

[13] `http://dict.leo.org`, eingesehen im Okt. 2006

gel sowohl in ihrer Muttersprache als auch in Englisch ein und führen die beiden Subwörter in einer Äquivalenzklasse zusammen. Neu erstellte Äquivalenzklassen werden daraufhin an Lexikographen mit anderer Muttersprache weitergeleitet, die passende Subwörter in ihrer Sprache ergänzen. Im Falle eines mehrdeutigen Subwortes wird die dazugehörige Äquivalenzklasse mittels *senseOf*-Relationen mit mindestens zwei nicht ambigen Äquivalenzklassen verknüpft. Bezüglich der *expandsTo*-Relation muss bisweilen entschieden werden, ob ein Mehrwort-Synonym anstelle einer *expandsTo*-Relation eingesetzt wird. Der Begriff *"Vitamin C"* beispielsweise kann durch folgende Arten mit *"ascorb"* in Beziehung gesetzt werden:

1. *Mehrwort-Subwort:* $(ascorb, ST, DE, d), (vitamin\ c, IV, DE, d) \in \mathcal{LE}_{\#ascorb}$
2. *expandsTo:* $(\#ascorb, (ascorb, ST, DE, CM),$
   $((\#ascorb, \#vitamin), (\#ascorb, \#letterc)), \varepsilon\} \in \mathcal{T}$

Das Vorgehen in letzterer Variante wird bevorzugt, wenn die einzelnen Wörter des Mehrwortausdruckes semantisch relevant sind, die erste Variante dementsprechend, wenn die einzelnen Wörter semantisch "schwach" sind. In diesem Falle wird wegen der semantischen Schwäche von *"C"* die erste Variante bevorzugt. Für weitere Beispiele zur Verwendung der *expandsTo*-Relation siehe Kapitel 2.3.

Eigennamen werden in die Lexika aufgenommen, wenn sie für die betrachtete Domäne relevant sind (wie in der medizinischen Domäne *"Crohn", "Parkinson"*). Auch hier werden Synonyme in gemeinsame Äquivalenzklassen zusammengefasst, was insbesondere bei pharmakologischen Produktnamen häufig auftritt, wenn Handelsnamen mit ihren Wirkstoffen verknüpft werden sollen(wie *Diclofenac, Voltaren,* ...).

2.5.3 Lexikon- und Thesaurus-Editor

Für die Arbeit an den Lexika und dem Thesaurus steht ein leistungsfähiges Editierwerkzeug zur Verfügung. Es ist Web-basiert und erlaubt parallele Lexikonarbeit an unterschiedlichen Orten. Die Oberfläche ist vertikal aufgeteilt und besteht im Kern aus zwei Hauptfenstern, die verschiedene Sichten auf die Einträge in den Lexika ermöglichen. Dies dient zur schnellen Verknüpfung von Subwörtern in gemeinsame Äquivalenzklassen oder zur Erzeugung von *expandsTo*- und *hasSense*-Relationen. Zahlreiche unterschiedliche Sortier- und Einschränkungsmöglichkeiten stehen beim Browsen durch die Lexika zur Verfügung. Zusätzlich wird der Zugriff auf mehrere Wortlisten ermöglicht, die bei der Integration von Subwörtern hilfreich sind. So kann sich ein Lexikograph alle Wörter und deren Worthäufigkeiten anschauen, die

Tabelle 2.3: Die lexikalischen Ressourcen von MorphoSaurus in Zahlen. $Subwörter_{all}$ - Gesamtanzahl der Subwörter, $Subwörter_{auto}$ - Anzahl der automatisch trainierten Subwörter, $EqKlassen$ - Anzahl der Äquivalenzklassen $Fract_{SubInEQ}$ - Verhältnis zwischen Subwörtern und Äquivalenzklassen ($Subwörter_{all}/EqKlassen$) eT - Anzahl der $expandsTo$-Relationen, sO - Anzahl der $senseOf$-Relationen.

Sprache	$Sub-wörter_{all}$	$Sub-wörter_{auto}$	$EqKlassen$	$Fract_{SubInEQ}$	eT	sO
EN	22.953	-	16.705	1,37	504	995
DE	24.357	-	16.750	1,45	590	1.019
PT	15.158	-	10.589	1,43	551	750
ES	13.390	5.641	9.593	1,40	427	583
FR	9.924	3.063	6.736	1,47	328	796
SE	13.981	5.694	8.045	1,74	348	1.191
All	99.781	14.389	22.707	4,39	1.108	3.332

ein bestimmtes Teilwort enthalten, was sich bei der Aufnahme kurzer Subwörter als nützlich erweist (siehe Kapitel 2.5.1). Abbildung 2.3 zeigt einen Screenshot des Editors.

2.5.4 Die lexikalischen Ressourcen in Zahlen

Die Erstellung der lexikalischen Ressourcen wurde in den letzten sechs Jahren mit unterschiedlichem Arbeitsaufwand bewerkstelligt. Das englische, deutsche und portugiesische Lexikon wurde in manueller Arbeit erstellt. Für Spanisch, Französisch und Schwedisch wurden Verfahren zur automatischen Erstellung und Erweiterung der lexikalischen Ressourcen entwickelt [Markó et al.2005c]. Die wichtigsten Kennzahlen der lexikalischen Ressourcen sind in Tabelle 2.3 aufgeführt.

Insgesamt enthält das Lexikon 99.781 Einträge[14] und 22.707 Äquivalenzklassen, was einem Verhältnis $Fract_{SubInEQ}$ von 4,39 entspricht. Betrachtet man nur deutsche Subwörter, so finden sich 24.357 Einträge und 16.750 Äquivalenzklassen. Dies sind im Schnitt 1,45 Subwort-Einträge pro Äquivalenzklasse.

Ein besonderer Vorteil bei der Verwendung von Subwörtern anstelle von ganzen

[14]Stand: Februar 2007

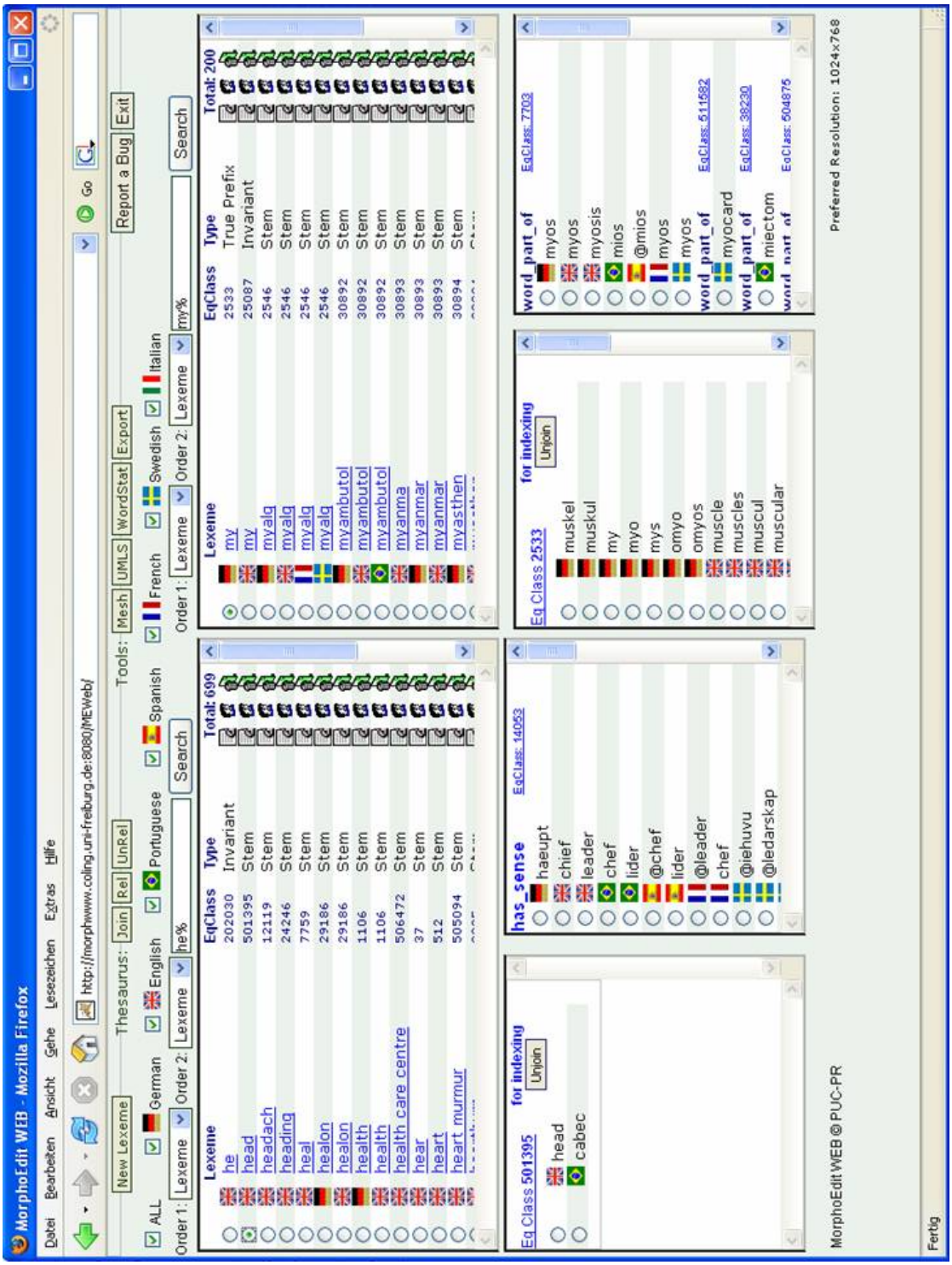

Abbildung 2.3: Screenshot des Lexikon-Editierwerkzeuges.

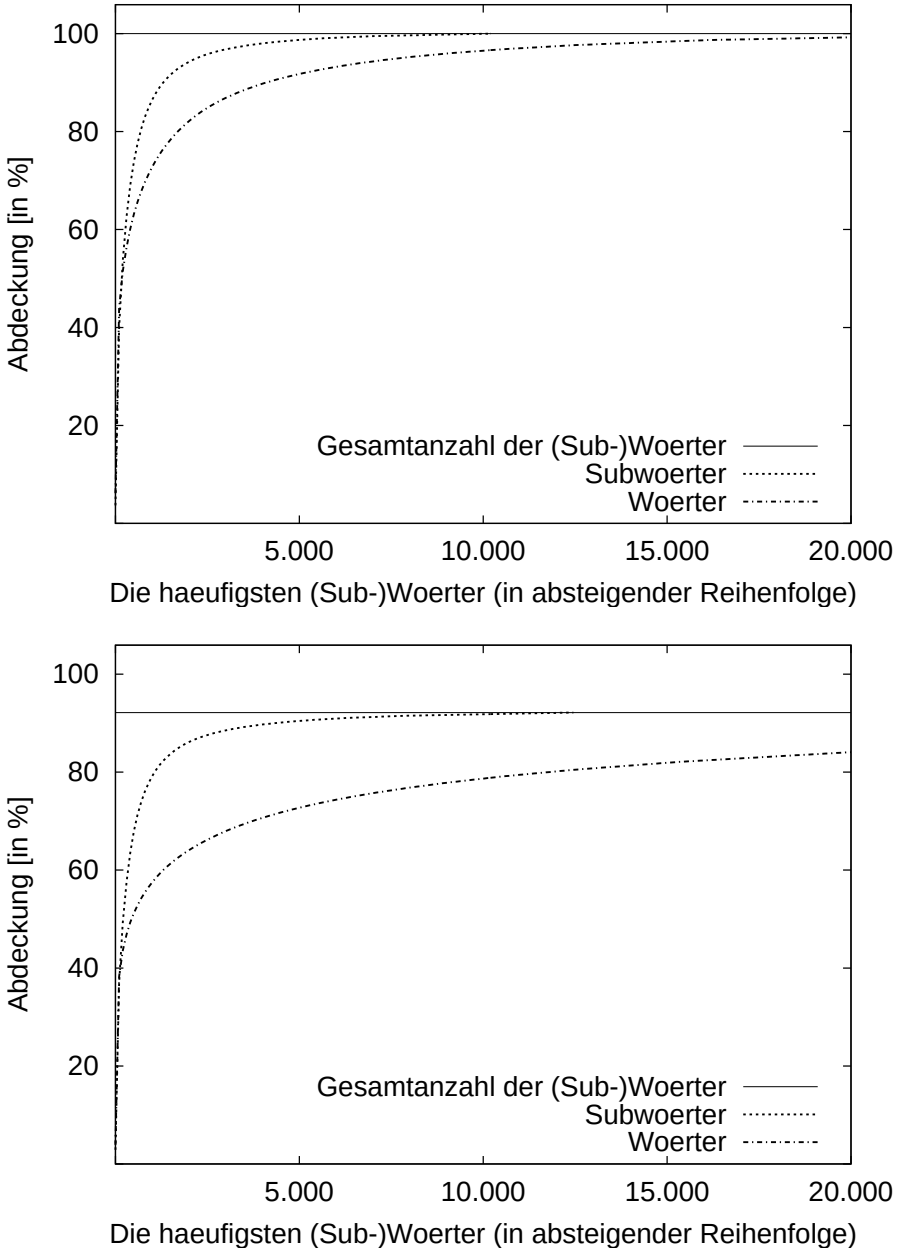

Abbildung 2.4: Vergleich Abdeckungsgrade von Dokumentenkollektionen zwischen Vollform-Wörter und Subwörter bei verschiedenen Cut-Off-Punkten.

Wörtern als Lexikoneinträge entsteht durch die deutliche Verringerung der Lexikoneinträge, die zur Abdeckung einer bestimmten Domäne benötigt werden. Durch den Verzicht der Aufnahme von Wortkomposita und -derivationen in das Lexikon bleibt die Anzahl der Einträge in den Lexika überschaubar. Dieses Phänomen wird anhand eines deutschen und englischen Textkorpus deutlich gemacht, deren Wörter eine Untermenge des Gesamtwortschatzes der medizinischen Domäne repräsentieren sollen. Es wurde bestimmt, aus wie vielen verschiedenen Wörtern diese Korpora zum einen in der Original-Version, zum anderen in der Subwort-Version zusammengesetzt sind. Außerdem wurde ermittelt, wie viele verschiedene Wörter benötigt werden, um einen gewissen Prozentanteil des Korpus abzudecken.

Insgesamt wurden 8082 MEDLINE-Abstracts in deutscher und englischer Sprache mit einer Gesamtanzahl von 1,4 Mio. englischen und 1,3 Mio. deutschen Wörtern akquiriert. Die Anzahl der unterschiedlichen Wörter in diesen Korpora betrug für das englische Korpus 25.478 Wörter und für das deutsche Korpus 66.964 Wörter.

Beginnend mit den am häufigsten auftretenden Wörtern wurde nun bestimmt, wieviel Prozent des Korpus aus diesen Wörtern besteht (siehe Tabelle 2.4). Betrachtet man die häufigsten 5.600 Wörter des deutschen Korpus, so werden 80% des deutschen Korpus aus diesen Wörtern gebildet. Mit 17.400 Wörtern können schon 90% des deutschen Korpus abgedeckt werden. Die Abbildung 2.4 verdeutlicht das typische asymptotische Verhalten dieser Zuordnungen.

Dieselben Prozentwerte wurden nun bestimmt, nachdem die Korpora in die Subwort-Version überführt wurden. Zur Erinnerung sei erwähnt, dass Wörter, die von MorphoSaurus nicht zerlegt werden können, in der Originalversion ausgegeben werden. Die Anzahl der Elemente eines Textes kann also höher sein als die Anzahl der Einträge in den MorphoSaurus-Lexika, wenn viele Subwörter in den Texten enthalten sind und zusätzlich einige Wörter nicht erkannt werden können. Insgesamt wurden 1,2 Mio. englische und 1,3 Mio. deutsche Subwörter erzeugt, die sich aus 10.204 verschiedenen englischen und 12.450 verschiedenen deutschen Subwörtern zusammensetzen. Bemerkenswert ist zunächst, dass die Anzahl der Subwörter in der morpho-semantisch normalisierten Version geringer ist als die Anzahl der Wörter in der Originalversion. Offensichtlich ist der Effekt, dass Stoppwörter von MorphoSaurus entfernt werden, größer als der Effekt, dass durch Zerlegung von Wörtern in Subwörter die Anzahl der Elemente ansteigt. Tabelle 2.4 macht außerdem deutlich, dass auffallend weniger Subwörter benötigt werden, um gewisse Prozentanteile des Korpus zu bilden. Beispielsweise bilden die häufigsten 700 Subwörter schon 80% des deutschen Korpus, was einer Reduktion an lexikalischen Einheiten auf 12,5% entspricht. Mit 1.400 Subwörtern werden schon 90% des Korpus gebildet (Reduktion

Tabelle 2.4: Grad der Abdeckung von Dokumentenkollektionen durch Vollform-Wörter und Subwörter bei verschiedenen Cut-Off-Punkten.

Deutsch				
Wörter	Abdeckung	Ganze Wörter	Subwörter	Verhältnis
1.031.009	80%	5.600	700	.13
1.174.568	90%	17.400	1.400	.08
1.239.821	95%	34.000	2.500	.07
1.265.923	97%	47.100	3.700	.08
1.292.024	99%	60.100	7.000	.12
1.305.075	100%	66.964	12.450	.19
Englisch				
Wörter	Abdeckung	Ganze Wörter	Subwörter	Verhältnis
1.132.906	80%	1.700	700	.41
1.274.520	90%	4.100	1.300	.32
1.345.326	95%	7.800	2.300	.29
1.373.649	97%	11.000	3.200	.29
1.401.972	99%	18.400	5.600	.30
1.416.133	100%	25.478	10.204	.40

auf 8%). Im Englischen ist dieses Phänomen etwas moderater mit einer Reduktion an lexikalischen Einheiten auf 29% bis 41% im Vergleich zur Originalversion. Dies verdeutlicht zugleich, dass die Wortbildungsphänomene wie Wortkomposition und Derivation im Deutschen deutlich ausgeprägter sind als im Englischen.

2.6 Vor- und Nachteile des MorphoSaurus-Systems

Das MorphoSaurus-System basiert auf der Annahme, dass aus semantischer Sicht die Zerlegung der Sprache in einzelne Wörter nicht die richtige Granularität widerspiegelt. Wie schon in der Einleitung ausführlich erläutert, erscheint es in vielen Anwendungen der natürlichen Sprachverarbeitung sinnvoll, semantisch atomare Einheiten anstelle von Wörtern voneinander abzugrenzen, wie dies im MorphoSaurus-System durch die Einführung von Subwörtern geschieht. Dadurch können

viele der sprachlichen Variationen, die in der natürlichen Sprache auftreten, adäquat behandelt werden.

Besonders einsichtig ist die Verwendung von Subwörtern auf der Ebene der morphologischen Variationen, schließlich ähneln die im MorphoSaurus-System verwendeten Subwörter den klassischen Morphemen der Wortbildungslehre. Ähnlich wie bei Verfahren zur Stammformbildung wird durch die Verwendung von Subwörtern eine morphologische Normalisierung erreicht.

Das MorphoSaurus-System geht jedoch einige Schritte weiter. Zum einen ist das lexikonbasierte Verfahren zur Dekomposition zusammengesetzter Wörter, welches aufgrund der qualitativ hochwertigen lexikalischen Ressourcen nur sehr selten zu Fehlzerlegungen führt, ein besonderer Mehrwert gegenüber regelbasierten Dekompositionsverfahren. Gegenüber Lexikon-basierten Dekompositionsverfahren, die auf Vollformen-Lexika aufbauen, ist ein entscheidender Vorteil, dass sich die wesentlichen Informationen von Dokumenten durch Reduktion auf Subwörter auf ein wesentlich geringeres Inventar an sprachlichen Einheiten reduzieren lässt, was den Pflegeaufwand der lexikalischen Ressourcen minimiert.

Ein weiterer Vorteil von MorphoSaurus gegenüber anderen Stammform- und Dekompositionsverfahren ergibt sich aus der Tatsache, dass Subwörter bzw. Äquivalenzklassen miteinander über verschiedene Relationen in Beziehung gebracht werden können. Somit werden die zwei für das Information Retrieval wichtigsten lexiko-semantischen Variationen, die Ambiguität einzelner Wörter einerseits sowie die Synonymie verschiedener Wörter andererseits, durch Äquivalenzklassen und *hasSense*-Relation umfassend behandelt. Dies ist in allen anderen dem Autor bekannten Verfahren nicht gegeben. Bezüglich der Synonymie muss allerdings festgestellt werden, dass eine semantische Unschärfe bei der Definition der Äquivalenzklassen zu deutlichen Einschränkungen der Performanz der auf dem MorphoSaurus-System aufbauenden Applikationen führen kann. Beispielsweise ist der Begriff *"somnolent"* inhaltlich verwandt mit dem Begriff *"schläfrig"*. *"Schläfrig"* wiederum kann von dem Begriff *"Schlaf"* abgeleitet werden. Die Eingruppierung der Begriffe *"Somnolenz", "somnolent", "schläfrig", "Schlaf"* in eine Synonymieklasse führt nun dazu, dass bei der Suche nach *"Somnolenz"* alle Dokumente, die in irgendeiner Form von *"Schlaf"* handeln, als gleich relevant eingestuft werden. Dies ist der Qualität eines IR-Systems sicherlich abträglich.

Auf die Aufnahme von Mehrwort-Einträgen in die MorphoSaurus-Subwort-Lexika wurde lange Zeit verzichet. Mittlerweile sind Mehrwort-Einträge in die Lexika erlaubt, um die wichtigsten Konzepte abzubilden, die durch Mehrwortausdrücke benannt sind. Der Eintrag von *"multiple sklerose"* in das deutsche Subwort-

Lexikon hat zur Folge, dass diesem Krankheitsbild mit *"#multiple_sklerose"* eine eigene Äquivalenzklasse zugeordnet wird. Nicht relevante Textpassagen wie "... *multiple* Läsionen mit deutlicher *Sklerose* der Kortikalis ..." werden somit in andere Äquivalenzklassen übersetzt und von einem IR-System nicht mehr als relevant eingestuft. Allerdings sind bei der Synonymie von Mehrwortausdrücken zwei Probleme zu berücksichtigen. Zum einen führt die Aufnahme von Mehrwortausdrücken zu erheblichem Mehraufwand bei der Pflege und Erweiterung der Subwort-Lexika. So ist es sehr aufwändig, die synonymen Ausdrücke *"postprandiale abdominale Beschwerden"* und *"epigastrische Schmerzen nach Nahrungsaufnahme"* adäquat miteinander zu verlinken, wenn man gleichzeitig die verschiedenen Variationen der beiden Ausdrücke wie Singular- und Pluralform mit berücksichtigen will. Zum anderen wird die Geschwindigkeit des MorphoSaurus-Systems durch die Suche nach Mehrworteinträgen negativ beeinträchtigt. Können die Dokumente bei Nichtberücksichtigung von Mehrwortausdrücken Wort für Wort in Subwörter zerlegt werden, so müssen nun Wörter einer bestimmten Fenstergröße auf entsprechende Subwort-Einträge überprüft werden. Da Geschwindigkeit bei großen Dokumentenkollektionen einen entscheidenden Parameter für die Qualität einer rechnergestützten Anwendung darstellt, ist bisher nur die Verwendung von Zweiwort-Ausdrücken abschließend unterstützt. Die Verwendung von Drei- und Mehrworteinträgen ist bisher nur experimentell implementiert.

Eine Voraussetzung für eine gute Performanz der auf dem MorphoSaurus-System aufbauenden Anwendungen ist die umfassende Abdeckung der behandelten Domäne in den Subwort-Lexika. Zwar werden Wörter, die vom MorphoSaurus-System nicht in Subwörter zerlegt werden können, in ihrer Originalform wieder zurückgegeben, so dass sie den Anwendungen weiter zur Verfügung stehen. Allerdings kann das Fehlen von Subwörtern durchaus auch Nachteile mit sich bringen, wie das Fehlen des Eintrags *"venedig"* im deutschen Subwort-Lexikon verdeutlicht. Das Fehlen führt dazu, dass das Wort in die für das MorphoSaurus-System korrekte Form *"vene⊕d⊕ig"* zerlegt wird. Die Suffixe *"d"* und *"ig"* werden als nicht relevant eingestuft, so dass für sie keine Äquivalenzklassen vorhanden sind. Somit wird das Subwort *"venedig"* in die Äquivalenzklasse *"#vene"* überführt. Dies hat in einem IR-System beispielsweise zur Folge, dass bei einer Anfrage nach *"Venen"* sämtliche Dokumente über *"Venedig"* angezeigt werden, was bei entsprechend großen Dokumentenkollektionen auch in der Medizin durchaus zu einigen Treffern führen kann. Eine ausreichende Abdeckung der MorphoSaurus-Lexika ist somit Grundvoraussetzung für das Funktionieren der auf MorphoSaurus aufbauenden Applikationen. Diese Abdeckung ist jedoch nur durch aufwändige manuelle Expertenarbeit

zu erreichen, was zeit- und kostenintensiv ist. Dies ist einer der Hauptnachteile des MorphoSaurus-Systems.

Das Problem der Erkennung von *"Venedig"* als Name einer Stadt könnte auch durch eine Anwendung zur Erkennung von Eigennamen (*Named Entity Recognition, NER*) gelöst werden. *NER* ist ein wichtiger Schritt zur semantischen Anreicherung von Texten. In der Medizin erlauben extrahierte Namen den gezielten Zugriff auf verschiedenste Stoff- und Objektklassen, wie zum Beispiel Namen von Organismen, Gewebs- und Zelltypen, Krankheiten, Medikamenten, Chemikalien, Proteinen oder Genen. Die Aufgabe eines automatisierten Systems zur Objekterkennung in Texten besteht darin, sämtliche Entitäten aufzuspüren und mit der entsprechenden Klassenbezeichnung zu markieren. So sollen unter anderem alle in einem Text auftretenden Medikamentnamen das Label "Medikament" erhalten. Eine zusätzliche Schwierigkeit ergibt sich aus der Verwendung linguistischer Variationen wie Synonymen, Abkürzungen und unterschiedlichen Schreibweisen für dieselbe Entität. Die Integration eines NER-Systems in MorphoSaurus ist geplant.

Durch die Normalisierung des MorphoSaurus-Systems gehen morphosyntaktische Informationen wie Wortart, Kasus, Numerus, Genus, Tempus oder Modus verloren. Die Verwendung von *Wortartenerkennung* oder *Chunking*, bei dem einzelne Wortformen zu einfachen Phrasen, etwa zu nicht-rekursiven Nominalphrasen, Verbalphrasen oder Präpositionalphrasen gruppiert werden, ist bisher nicht in das MorphoSaurus-System integriert. Es ist daher verständlich, dass das MorphoSaurus in erster Linie für Anwendungen geeignet ist, bei dem syntaktische Informationen eine nur untergeordnete Rolle spielen. Dies ist im Information Retrieval in besonderem Maße gegeben. Hier sind die Erfolge, die durch Verfahren wie *Wortartenerkennung* oder *Chunking* erzielt werden, insgesamt nur moderat[Hersh2002]. Außerdem werden durch eine wortbasierte Suchmaschine wie die von uns verwendete Open-Source Software *Lucene* (siehe Kapitel 3.3) einige der syntaktischen Variationen bereits berücksichtigt. Dies liegt daran, dass die Wortreihenfolge bei dieser Suchmaschine keine Rolle spielt. Beispielsweise werden in den Ausdrücken *"Sein Blutdruck war bei Aufnahme deutlich erhöht"* und *"Bluthochdruck"* unabhängig von der Reihenfolge der Wörter die gleiche Anzahl relevanter Subwörter gefunden, wenn nach *"Hypertonie"* gesucht wird, nämlich *"#high", "#blood"* und *"#pressure"*. Lediglich in Situationen, in denen die Wortreihenfolge entscheidend zur Semantik eines Ausdruckes beiträgt, ist der nicht phrasenbasierte Suchansatz hinderlich, wie die zwei Ausdrücke *"Expertensysteme zur Verbesserung medizinischer Diagnosen"* und *"medizinische Diagnosen zur Verbesserung von Expertensystemen"* verdeutlichen, deren Bedeutung durch die veränderte Wortstellung der einzelnen Wörter unter-

schiedlich ist. Für andere Anwendungen der rechnergestützten natürlichen Sprachverarbeitung, wie der maschinellen Übersetzung oder der automatischen Textzusammenfassung, erscheint die Anwendung von MorphoSaurus nur eingeschränkt oder in Kombination mit weiteren Werkzeugen sinnvoll, da hier die Verwendung morpho-syntaktischer Informationen von entscheidender Bedeutung ist.

2.7 Verwandte Arbeiten

Dem Autor ist kein weiteres aktuelles Forschungsprojekt bekannt, welches sich mit der Erstellung von mehrsprachigen lexikalischen Ressourcen auf Morphem- oder Subwort-Ebene beschäftigt. Es gibt jedoch eine Reihe allgemeiner und biomedizinischer Vokabulare auf der Ebene von Wort- und Mehrwort-Einträgen, die Gemeinsamkeiten und Unterschiede zu unseren Ressourcen aufweisen und an dieser Stelle diskutiert werden sollen. Verwandte Arbeiten über Algorithmen, die eine zu MorphoSaurus vergleichbare sprachliche Normalisierung durchführen, wurden bereits in der Einleitung diskutiert oder werden noch einmal im Kontext des Information Retrievals in Kapitel 3.7 aufgegriffen.

Das bekannteste Beispiel lexikalischer Ressourcen in der Biomedizin ist das Unified Medical Language System ([Umls2005b]), welches an der U.S. Library of Medicine entwickelt und gepflegt wird. Der Metathesaurus enthält über einhundert unterschiedliche medizinische Terminologien mit mehr als 1,3 Millionen Konzepten und 6,4 Millionen Konzeptbezeichnern, viele hiervon in mehreren Sprachen. Ein UMLS-*Konzept* ist eine Gruppe synonymer Konzeptbezeichner und kann mit den Äquivalenzklassen in MorphoSaurus verglichen werden kann. Bezüglich lexikalischer (Term-Term) und semantischer (Konzept-Konzept) Relationen übernimmt UMLS die Beziehungen der zugrunde liegenden Vokabularien, so dass eine Gesamtzahl von über 50 Relationen resultiert. Verglichen mit dem schlanken Relationen-System in MorphoSaurus, in dem lediglich *hasSense* und *expandsTo*-Relationen definiert sind, ist UMLS also wesentlich umfangreicher und enthält eine Vielzahl hierarchischer und nicht-hierarchischer Beziehungen wie *Is_A*, *Part_Of* und *Consists_Of*-Relationen.

Die "Systematized Nomencalture of Medincine - Clinical Terms" [Snomed CT2006] ist als eine umfassende Referenzterminologie für alle Bereiche des Gesundheitswesen entwickelt worden und soll in den nächsten Jahren in verschiedenen Ländern routinemäßig eingesetzt werden. Sie erhebt den Anspruch einer formalen, auf Beschreibungslogik basierenden Semantik. Ein Teil der SNOMED CT Terminologie ist auch im UMLS enthalten. Das Vokabular ist in unterschiedli-

chem Umfang in verschiedene andere Sprachen übersetzt. Es ist ein konzeptbasiertes kontrolliertes Vokabular, welches als grundlegende Elemente *Konzepte, Hierarchien, Relationen* und *Beschreibungen* enthält. Konzepte bestehen aus einem numerischen Code, einem einmaligen Namen sowie Beschreibungen, welche einen bevorzugten Term und eine oder mehrere Synonyme enthalten. Sie entsprechen am ehesten den Äquivalenzklassen des MorphoSaurus-Systems, wobei für Äquivalenzklassen beispielsweise keine Vorzugsterme definiert sind. Die Beschreibungen entsprechen dann grob den Subwörtern von MorphoSaurus. Derzeit sind in der englischen Version von SNOMED CT 308.000 Konzepte und 777.000 Beschreibungen definiert[15]. Sie sind derzeit in 19 Hauptkategorien sowie weiterer Unterkategorien eingebettet. Über semantische Relationen, wie z.B. *"has-associated-topography", "has-action", "has-associated-morphology"* sind die Konzepte miteinander verknüpft. Mit Hilfe der SNOMED-Terminologie lassen sich unter anderem Befunde, Diagnosen oder Therapien sehr detailliert beschreiben. Dabei können zwei Vorgehensweisen unterschieden werden, die beide von SNOMED unterstützt werden. Einerseits werden Konzepte vorab definiert und im System bereitgestellt. Dieses in SNOMED CT vorherrschende Prinzip der Vorab-Kombination von Konzepten (Präkoordination) führt zu der sehr großen Zahl von Einträgen. Die alternative Vorgehensweise der Kombination von Konzepten zum Verwendungszeitpunkt (Postkoordination) wird möglichst beschränkt auf ergänzende Angaben (Qualifier) wie Seitigkeit (rechts, links) und Verlauf (akut, chronisch). SNOMED ist die umfassendste Terminologie ihrer Art, jedoch zeigen sich gerade in der komplexen Anordnung von Relationen auch Schwächen dieser Terminologie [Schulz et al.2005].

Der MeSH (Medical Subject Headings)[MeSH2006] ist ein Vokabular, das ursprünglich für die Kategorisierung der in der MEDLINE-Literaturdatenbank verfügbaren Artikel entwickelt und mittlerweile in andere Sprachen übersetzt wurde. Er wird ausführlich in Kapitel 4.2 vorgestellt.

WordNet[Fellbaum1998] ist das bekannteste nicht Domänen spezifische System, welches an der Universität von Princeton entwickelt wurde. Zusätzlich existiert das europäische EuroWordNet[Vossen1998], welches verschiedene europäische Sprachen enthält, wobei die terminologische Abdeckung in den unterschiedlichen Sprachen sehr variiert. WordNet besteht aus kanonischen Formen von Termen und enthält derzeit ungefähr 300.000 Einträge (einschließlich EuroWordNet). Eine Gruppe synonymer Terme in WordNet wird *Synset* genannt, welche mit den UMLS-Konzepten und den MorphoSaurus-Äquivalenzklassen vergleichbar sind. Im Un-

[15]Stand: Januar 2007

terschied zum MorphoSaurus-System, in dem für doppeldeutige Subwörter eigene Äquivalenzklassen definiert sind, können doppeldeutige Terme in WordNet in mehreren Synsets auftreten[16]. WordNet definiert außerdem die sechs Beziehungstypen *Hyponymie, Synonymie, Meronymie, Antonymie, Troponomie* und *Implikation*.

Allen vorgestellten Terminologien ist gemeinsam, dass sie sich in Art und Umfang der Einträge und der Relationen deutlich vom MorphoSaurus-System unterscheiden. Durch die Verwendung von Wörtern und Mehrworteinträgen als lexikalische Einheiten sind in den diskutierten Terminologien wesentlich mehr Begriffe enthalten als in den MorphoSaurus-Lexika. Das Vorhandensein dieser Terminologien verdeutlicht, dass es nicht Zweck und Ziel des MorphoSaurus-Systems sein kann, parallele Lexika, die die Funktion dieser Terminologien imitieren, eigenständig aufzubauen. Vielmehr sollte versucht werden, die lexikalischen Ressourcen von MorphoSaurus auf andere bestehende Terminologien abzubilden beziehungsweise vorhandene Synergien zu nutzen.

[16]Der Term *"bank"* beispielsweise erscheint in 18 verschiedenen Synsets

Kapitel 3

Dokumentenrecherche mit MorphoSaurus

3.1 Einleitung

Die Evaluation von IR-Systemen hat in den letzten Jahren durch verschiedene Evaluationskampagnen wie die Text Retrieval Conference (TREC)[1] oder das Cross Lingual Evaluation Forum (CLEF)[2] enormen Aufschwung erlebt. Das Ziel solcher Evaluationskampagnen ist es, Forscher und ihre Systeme zusammenzubringen, eine wohl definierte, realistische Testbasis zur Verfügung zu stellen, und den Vergleich unterschiedlicher Methoden und Ideen zu ermöglichen. Forscher sollen ihre neue Entwicklungen in realistischen Umgebungen auf die Bedürfnisse der Wirtschaft und zukünftigen Benutzer anpassen. Der Vergleich erfolgt durch Verwendung gemeinsamer Testanfragen und -daten, für die fachbezogene Relevanzbeurteilungen bezüglich der Testanfragen manuell oder halbautomatisch durchgeführt werden. Auch das MorphoSaurus-System nahm im Jahr 2006 mit Erfolg an der CLEF-Kampagne im Track "Cross-Language Retrieval in Image Collections (ImageCLEF)" teil [Daumke & Markó2006].

In den vorangegangen Artikeln wurde MorphoSaurus als ein System vorgestellt, welches angemessene Lösungen für zahlreiche linguistische Herausforderungen bietet und beispielsweise im Information Retrieval eingesetzt werden kann. In diesem Kapitel soll der Nutzen von MorphoSaurus in der einsprachigen Dokumentenrecherche anhand verschiedener Testkollektionen evaluiert werden, die auch in den Evaluationskampagnen verwendet werden. Die mehrsprachige Recherche wird in dieser Arbeit nicht betrachtet, für Informationen hierzu wird auf [Markó2007] verwie-

[1] `http://trec.nist.gov/`, eingesehen im Februar 2007

[2] `http://www.clef-campaign.org/`, eingesehen im Februar 2007

sen. Da das MorphoSaurus-System auf die biomedizinische Domäne spezialisiert ist, werden zwei medizinische Dokumentenkollektionen eingesetzt. Darüber hinaus wird das MorphoSaurus auch auf einem sozialwissenschaftlichen Korpus getestet, um die Nützlichkeit des MorphoSaurus-Systems in einer nicht-medizinischen Domäne zu evaluieren. Die Dokumentenkollektionen, die in diesem Evaluationsszenario zum Einsatz kommen, sind im Einzelnen:

- **OHSUMED**: Die *OHSUMED*-Kollektion ist eine im Jahr 1994 von William Hersh erstellte englischsprachige Testdatenbank für medizinische Recherchesysteme, stellt eine Untermenge von MEDLINE dar und wurde unter anderem im TREC-9 Filtering Track[Robertson & Soboroff2001] eingesetzt.

- **ImageCLEFmed**: Die *ImageCLEFmed 2006*-Kollektion ist eine Sammlung aus vier verschiedenen Datenquellen (*Casimage, Peir, MIR, PathoPic9*) [Müller et al.2006a], die in den Jahren 2005 und 2006 in dem Track "Cross-Language Retrieval in Image Collections" verwendet wurden. Die enthaltenen Sprachen für die Bildannotationen beinhalten Englisch, Deutsch und Französisch. Bei *ImageCLEF* können sowohl textbasierte als auch visuelle Recherchesysteme teilnehmen. Diese Experimente konzentrieren sich auf die textuelle Suche von Bildern anhand der Bildunterschriften.

- **GIRT**: Die *GIRT*-Kollektion ist eine sozialwissenschaftliche Test-Datenbank, die seit dem Jahr 2000 in der domänenspezifischen Aufgabe der CLEF-Kampagne eingesetzt wird. Sie enthält Daten aus den Datenbanken *FORIS* und *SOLIS* [Kluck2004], die dem Datenbestand des Informationszentrums Sozialwissenschaften entnommen wurden. Sie liegen in deutscher und englischer Sprache vor.

Bei der Evaluation sollen verschiedene Ansätze der morphologischen Dokumentenanalyse miteinander verglichen werden, insbesondere werden die Effekte regelbasierter Stammformbildung mit dem Porter-Algorithmus ([Porter1980]) sowie der morpho-semantischen Normalisierung mit der MorphoSaurus-Technologie untersucht.

Zusätzlich zu der linguistischen Vorverarbeitung durch den Porter-Stemmer sowie durch das MorphoSaurus-System wird eine Suchmaschine benötigt, die die eigentliche Suche basierend auf den Originalwörtern bzw. auf den vorverarbeiteten Texten durchführt. In dieser Arbeit kommt die Open-Source-Software *Lucene* der Apache Foundation[3] zum Einsatz.

[3] `http://jakarta.apache.org/lucene/docs/index.html`, eingesehen im Okt. 2006

In den folgenden Abschnitten werden zunächst die Testkollektionen näher beschrieben sowie einige Charakteristika der Suchmaschine *Lucene* vorgestellt. Anschließend werden die Durchführung und die Evaluation der Experimente erläutert.

3.2 Die Test-Kollektionen

3.2.1 OHSUMED-Kollektion

OHSUMED-Dokumente

Die OHSUMED-Testkollektion[Hersh et al.1994a, Hersh & Dickham1994] besteht aus einer Stichprobe von 348.566 Referenzen aus MEDLINE, der weltweit größten medizinischen Datenbank, die von der National Library of Medicine erstellt wird. Sie wurde unter anderem im TREC-9 Filtering Track[Robertson & Soboroff2001] eingesetzt. Die Referenzen enthalten Titel und/oder Abstracts aus 270 medizinischen Zeitschriften über einen Zeitraum von fünf Jahren (1987-1991). Titel und Abstract bestehen insgesamt aus 26.705.691 Wörtern, was einer durchschnittlichen Dokumentenlänge von 76,6 Wörtern entspricht. Neben Titel und Abstract besteht eine Referenz aus MeSH-Schlagwörtern, Autoren, Quellenangaben sowie aus dem Typ der Veröffentlichung. Listing 3.1 zeigt ein typisches Dokument aus der OHSUMED-Kollektion. Im IR-Szenario wurden der Titel, das Abstract und die MeSH-Schlagwörter berücksichtigt (siehe Tabelle 3.2).

OHSUMED-Testanfragen

Insgesamt stehen 106 Testanfragen zur Verfügung, für welche die Relevanz der OHSUMED-Dokumente manuell von Experten der Oregon Health Science University[Hersh et al.1994a] beurteilt wurde. Anfragen wurden einem Online-Fragebogen entnommen, welcher einem Suchinterface für MEDLINE vorgeschaltet wurde. Anfragen bestehen aus Titel und Beschreibung, wobei der Titel eine Patientenbeschreibung enthält und die Beschreibung das eigentliche Informationsbedürfnis des Benutzers formuliert. Die durchschnittliche Länge der Beschreibung, die als Anfrage für diese Experimente verwendet wurde, beträgt 6,77 Wörter pro Anfrage. Um die Relevanzbeurteilungen der OHSUMED-Dokumente bezüglich der Testanfragen zu erstellen, führten jeweils zwei Ärzte und zwei Bibliothekare anhand des Titels und der Beschreibung Suchanfragen in MEDLINE aus, wobei mehrfache Anfragen möglich waren und sämtliche in MEDLINE verfügbaren Suchmöglichkeiten verwendet werden konnten. Anschließend wurden alle MEDLINE-Treffer, die nicht

Listing 3.1: Typische Referenz aus der OHSUMED-Kollektion.

```
<RECORD>
 <ID>1</ID>
 <MEDLINE-ID>87049087</MEDLINE-ID>
 <SOURCE>Am J Emerg Med 8703; 4(6):491-5</SOURCE>
 <MESH>
  Allied Health Personnel/*;
  Electric Countershock/*;
  Emergencies;
  Emergency Medical Technicians/*;
  Human;
  Prognosis;
  Recurrence;
  Support, U.S. Gov't, P.H.S.;
  Time Factors;
  Transportation of Patients;
  Ventricular Fibrillation/*TH.
 </MESH>
 <TITLE>
  Refibrillation managed by EMT-Ds: incidence and outcome without paramedic back-up.
 </TITLE>
 <PTYPE>JOURNAL ARTICLE</PTYPE>
 <ABSTRACT>
  Some patients converted from ventricular fibrillation to organized rhythms by
  defibrillation-trained ambulance technicians (EMT-Ds) will refibrillate before
  hospital arrival. The authors analyzed 271 cases of ventricular fibrillation
  managed by EMT-Ds working without paramedic back-up. Of 111 patients initially
  converted to organized rhythms, 19 (17%) refibrillated, 11 (58%) of whom were
  reconverted to perfusing rhythms, including nine of 11 (82%) who had spontaneous
  pulses prior to refibrillation. Among patients initially converted to organized
  rhythms, hospital admission rates were lower for patients who refibrillated than
  for patients who did not (53% versus 76%, P = NS), although discharge rates
  were virtually identical (37% and 35%, respectively). Scene-to-hospital
  transport times were not predictively associated with either the frequency of
  refibrillation or patient outcome. Defibrillation-trained EMTs can effectively
  manage refibrillation with additional shocks and are not at a significant
  disadvantage when paramedic back-up is not available.
 </ABSTRACT>
 <AUTHOR>Stults KR; Brown DD.</AUTHOR>
</RECORD>
```

Listing 3.2: Typische Testanfrage aus der OHSUMED-Kollektion.

```
<title>
        60 year old menopausal woman without hormone replacement therapy
</title>
<desc>
        Are there adverse effects on lipids when progesterone is given with
        estrogen replacement therapy?
</desc>
```

in der OHSUMED-Kollektion enthalten waren, verworfen. Die restlichen Treffer wurden von Ärzten auf Relevanz bezüglich der Testanfragen beurteilt. Relevante Dokumente wurden als *"potentiell relevant"* oder *"definitiv relevant"* markiert. 8.714 Treffer wurden als nicht relevant, 2.053 als potentiell relevant und 1.798 als definitiv relevant beurteilt. 1.453 (11,4%) Dokumente wurden doppelt beurteilt, die Übereinstimmung zwischen zwei Personen lag bei 1.003 Übereinstimmungen und 432 Nichtübereinstimmungen (kappa-Wert von 0.41). Klassischerweise werden bei IR-Evaluationsszenarien sowohl potentiell relevante als auch definitiv relevante Markierungen als relevant angesehen. In Listing 3.2 ist eine typische Testanfrage aus der OHSUMED-Kollektion abgebildet.

3.2.2 ImageCLEF-Kollektion

ImageCLEF-Dokumente

Die ImageCLEF-Kollektion ist eine Testdatenbank für die Bildersuche, in der Bilder und Bildunterschriften von vier verschiedenen Datenquellen enthalten sind. Die Casimage-Datenbank[4] enthält circa 2.000 medizinische Fallbeispiele mit knapp 9.000 Bildern, darunter radiologische Bilder, Fotographien und PowerPoint-Folien [Müller et al.2004]. Die meisten Fälle sind in französischer Sprache geschrieben, 20% der Fälle liegen in Englisch vor und 5% der Bilder besitzen keine Annotation. PEIR[5] (Pathology Education Instructional Resource) ist aus dem HEAL-Projekt (Health Education Assets Library) heraus entstanden [Candler et al.2003]. Es enthält 33.000 Bilder vornehmlich aus der Pathologie, die mit englischen Annotationen versehen sind. Im Gegensatz zu Casimage, in der die Annotationen fallbasiert sind, gibt es in dieser Kollektion zu jedem Bild eine Annotation. Die nuklearmedizinische Datenbank von MIR, dem Mallinkrodt Institut für Radiologie[6] enthält 2.000 Bilder mit englischen, fallbasierten Annotation [Wallis et al.1995]. Die PathoPic-Sammlung[7] beinhaltet 9.000 Bilder mit ausführlicher Pro-Bild-Annotation in deutscher Sprache [Glatz-Krieger et al.2003]. Teile der deutschen Annotation sind ins Englische übersetzt worden. Insgesamt stehen in der ImageCLEF-Kollektion also ca. 50.000 Bilder aus vier verschiedenen Datenquellen mit Annotation in englischer, deutscher und französischer Sprache zur Verfügung. Abbildung 3.1 zeigt beispielhaft ein Dokument aus der PathoPic-Kollektion.

[4] `http://www.casimage.com`, eingesehen im Februar 2007
[5] `http://peir.path.uab.edu/`, eingesehen im Februar 2007
[6] `http://gamma.wustl.edu/home.html`, eingesehen im Februar 2007
[7] `http://alf3.urz.unibas.ch/pathopic/intro.htm`, eingesehen im Februar 2007

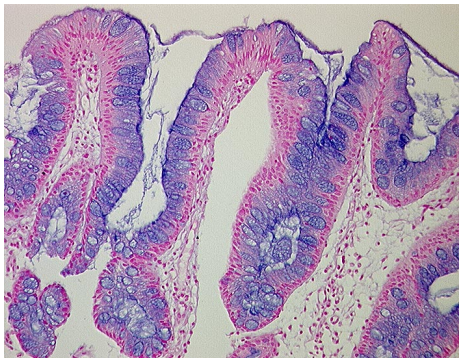

Abbildung 3.1: Beispiel für ein Dokument aus der ImageCLEF–Kollektion.
ID: 000001, Diagnose: Barrett Oesophagus, Synonyme: intestinale Metaplasie, Beschreibung: Zylinderepithelmetaplasie mit alcianblauen Schleim bildenden Becherzellen. Villöse Epitheloberflaeche. Epithel ohne dysplastische Veränderungen. Klinik: GERD mit retrosternalen Schmerzen.

Die Bildunterschriften zu allen vier Datenbanken liegen als XML-Dokumente vor. Jede Datenbank enthält unterschiedliche, teils zahlreiche XML-Tags. Eine Auflistung erscheint in diesem Rahmen nicht sinnvoll. Zusammenfassend lässt sich sagen, dass alle aus IR-Sicht relevanten textuellen Informationen aus den Dokumenten extrahiert wurden und in den Suchmaschinen-Index überführt wurden (siehe Tabelle 3.2 auf Seite 62).

ImageCLEF-Testanfragen

Die Testanfragen von ImageCLEFmed 2006 basieren auf zwei Erhebungen, die in Portland und Genf durchgeführt wurden [Hersh et al.2005, Müller et al.2006b]. Zusätzlich wurde eine Log-Datei des Multimediasuchdienstes von Health on the Net (HON) [8] ausgewertet. Auf Basis dieser Daten wurden Testanfragen ausgewählt, die eine oder mehrere der folgenden Merkmale aufwiesen:

- Angabe einer anatomischen Region.
- Angabe einer Modalität, durch welche das Bild erstellt wurde (Röntgen, CT, MRT, Ultraschall, etc.).
- Angabe eines pathologischen Befundes bzw. einer Erkrankung.
- Sonstige abnorme Beobachtungen (bspw. ein vergrößertes Herz).

[8] `http://www.hon.ch/`, eingesehen im Februar 2007

Die ausgewählten Fragen wurden dahingehend sortiert, ob sie eher durch ein visuelles, ein textuelles oder ein gemischtes IR-System auffindbar sind. Schließlich wurden für jede dieser drei Kategorien 10 Anfragen extrahiert, die mindestens eines der oben genannten Merkmale enthielt. Alle 30 Fragen wurden von Experten ins Deutsche und ins Französische übersetzt. Die durchschnittliche Anzahl der Wörter pro Frage beträgt im Englischen (Deutschen; Französischen) 5,73 (4,47;5,5) Wörter. Abbildung 3.2 zeigt eine typische Frage aus der ImageCLEFmed-Kampagne von 2006.

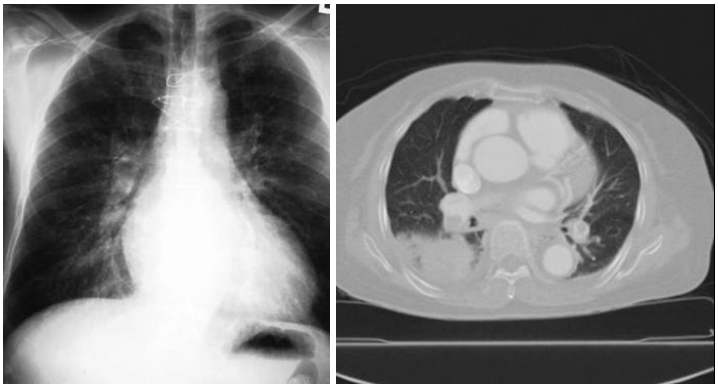

Abbildung 3.2: Beispiel fuer eine Frage in ImageCLEFmed 2006
"Show me CT or x-ray images showing the heart."
"Zeige mir CT Bilder oder Röntgenbilder des Herzens."
"Montre-moi des images CT ou des radiographies qui montrent le coeur."

Für die Relevanzbeurteilungen wurden Bildmengen zwischen 647 und 1.187 Bildern pro Anfrage erzeugt, mit durchschnittlich 910 Bildern pro Anfrage. Diese wurden von sieben Ärzten der Oregon Health and Science University auf Relevanz bezüglich der Testanfrage beurteilt.

3.2.3 GIRT-Kollektion

Girt-Dokumente

GIRT (German Indexing and Retrieval Test Database) ist eine sozialwissenschaftliche Test-Datenbank, welche eine Grundlage für eine objektive Bewertung und Vergleichbarkeit verschiedener Retrievalsysteme in der Sozialmedizin schaffen möchte [Kluck2004]. Erstellt und gepflegt wird sie vom Informationszentrum Sozialwissenschaften (IZ). Die vierte Version des GIRT-Korpus wurde unter anderem für das *Cross Language Evaluation Forum* (CLEF) in den Jahren 2003 bis 2006 verwendet und wird für CLEF 2007 angeboten.

Die GIRT-Daten bestehen aus Dokumenten, welche dem Datenbank-Bestand des Informationszentrums Sozialwissenschaften (IZ)[9] entnommen wurden. Als Datenbanken dienen FORIS[10] und SOLIS[11] (Sozialwissenschaftliches Literaturinformationssystem). Inhalte der FORIS-Datenbank sind Beschreibungen verschiedener Forschungsprojekte aus Deutschland, Schweiz, Österreich und Liechtenstein. Der Datenbankbestand enthält ca. 40.000 Projekte, wovon ca. 6.000 wegen Fortschreibung der Projekte jährlich geändert werden. SOLIS enthält Informationen über deutschsprachige sozialwissenschaftliche Literatur wie Dissertationen, Habilitationen, Monographien, Forschungsberichte oder Sammelwerke. Der Datenbankbestand enthält derzeit ca. 325.000 Dokumente bei einem jährlichen Zuwachs von 12.000 bis 14.000 Dokumenten. Damit stammen die GIRT-Daten aus sozialwissenschaftlichen Fachinformationen, die vom IZ aufbereitet und öffentlich gegen Entgelt angeboten werden. Sie sind im Original deutschsprachig, wobei der Titel, die Inhaltsangaben und die Deskriptoren der meisten Dokumente ins Englische übersetzt sind, um mehrsprachige Suche zu ermöglichen. Die GIRT-Daten enthalten in beiden Sprachen mindestens die folgenden Felder: Autor, deutscher Titel, Sprache des Dokumentes, Erscheinungsjahr sowie Schlagwörter und Klassifikationstexte.

Die GIRT-Daten liegen mittlerweile in der vierten Generation vor (GIRT4). Die Auswahl und der Umfang der Dokumente wurden für die vierte Generation grundlegend neu gewählt. GIRT4 liegt auf Deutsch (GIRT4-DE) und auf Englisch (GIRT4-EN) in zwei getrennten Korpora vor. Die Gesamtzahl der GIRT4-Dokumente beträgt in beiden Sprachen jeweils 151.319 Dokumente. Dabei muss in den englischen Dokumenten zumindest der Titel des Dokumentes aus dem Deutschen übersetzt sein. Die Erscheinungsjahre wurden auf 1990 bis 2000 begrenzt. Die Felder, die in GIRT4 verwendet werden, sind in Tabelle 3.1 auf Seite 60 dargestellt. In den Lucene-Index aufgenommen wurden Titel, Abstract, Schlagwörter und Klassifikationstexte (siehe Tabelle 3.2 auf Seite 62)).

Listing 3.3 bis 3.6 stellen ein typisches GIRT-Dokument in einer XML-Version (Listing 3), in einer von Stoppwörtern bereinigten Version (Listing 4), in einer gestemmten Version (Listing 4) sowie in einer MorphoSaurus-Version dar.

[9] `http://www.gesis.org/iz/`, eingesehen im Okt. 2006

[10] `http://www.gesis.org/Information/FORIS/Recherche/index.htm` (Forschungsinformationssystem Sozialwissenschaften) , eingesehen im Okt. 2006

[11] `http://www.gesis.org/Information/SOLIS/index.htm`, eingesehen im Okt. 2006

Listing 3.3: Typisches Dokument aus der GIRT-Dokumentenkollektion in XML-Darstellung. Aus Platzgründen wird nur der erste Satz des Abstracts angezeigt.

```
<DOC>
 <DOCNO>GIRT-DE19909343</DOCNO>
 <DOCID>GIRT-DE19909343</DOCID>
 <TITLE-DE>
        Die sozioökonomische Transformation einer Region:
        Das Bergische Land von 1930 bis 1960
 </TITLE-DE>
 <AUTHOR>Henne, Franz J.</AUTHOR>
 <AUTHOR>Geyer, Michael</AUTHOR>
 <PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
 <LANGUAGE-CODE>DE</LANGUAGE-CODE>
 <CONTROLLED-TERM-DE>Rheinland</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>historische Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>regionale Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>sozioökonomische Faktoren</CONTROLLED-TERM-DE>
 <METHOD-TERM-DE>historisch</METHOD-TERM-DE>
 <METHOD-TERM-DE>Aktenanalyse</METHOD-TERM-DE>
 <CLASSIFICATION-TEXT-DE>Sozialgeschichte</CLASSIFICATION-TEXT-DE>
 <ABSTRACT-DE>Die Arbeit hat das Ziel, anhand einer regionalen Studie
        die Entstehung des "modernen" fordistischen Wirtschaftssystems und
        des sozialen Systems im Zeitraum zwischen 1930 und 1960 zu beleuchten ...
 </ABSTRACT-DE>
</DOC>
```

Listing 3.4: Darstellung des Dokumentes ohne Stoppwörter. Angegeben sind diejenigen XML-Tags, deren textuellen Inhalte in der Indexierungsphase analysiert wurden.

```
...
 <TITLE-DE>
        sozioökonomische Transformation Region : Bergische Land 1930 1960
 </TITLE-DE>
...
 <CONTROLLED-TERM-DE>Rheinland</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>historische Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>regionale Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>sozioökonomische Faktoren</CONTROLLED-TERM-DE>
...
 <CLASSIFICATION-TEXT-DE>Sozialgeschichte</CLASSIFICATION-TEXT-DE>
 <ABSTRACT-DE>
        Arbeit Ziel anhand regionalen Studie Entstehung modernen fordistischen
        Wirtschaftssystems sozialen Systems Zeitraum 1930 1960 beleuchten ...
 </ABSTRACT-DE>
</DOC>
```

Listing 3.5: Darstellung der entsprechenden XML-Felder, nachdem deren Inhalte mit dem Porter-Stemmer verarbeitet wurden.

```
...
 <TITLE-DE>
        soziookonom transformation region : bergisch land 1930 1960.
 </TITLE-DE>
...
 <CONTROLLED-TERM-DE>Rheinland</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>histor Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>regional Entwicklung</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>soziookonom Faktor</CONTROLLED-TERM-DE>
...
 <CLASSIFICATION-TEXT-DE>#socio #histor</CLASSIFICATION-TEXT-DE>
 <ABSTRACT-DE>
        Arbeit Ziel anhand regional Studi Entstehung modern fordist
        Wirtschaftssystem sozial System Zeitraum 1930 1960 beleuchten ...
 </ABSTRACT-DE>
</DOC>
```

Listing 3.6: Die morpho-semantisch normalisierte Darstellung des MorphoSaurus-Systems. Das Wort “fordistischen”konnte nicht zerlegt werden, daher bleibt die ursprüngliche Darstellung des Wortes erhalten.

```
...
 <TITLE-DE>
        #socio #econom #transform #area : #mountain #country 1930 1960
 </TITLE-DE>
...
 <CONTROLLED-TERM-DE>#rhein #country</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>#histor #develop</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>#area #develop</CONTROLLED-TERM-DE>
 <CONTROLLED-TERM-DE>#socio #econom #factor</CONTROLLED-TERM-DE>
...
 <CLASSIFICATION-TEXT-DE>#socio #histor</CLASSIFICATION-TEXT-DE>
 <ABSTRACT-DE>
        #work #goal #area #study #develop #modern fordistischen
    #econom #system #socio #system #period 1930 1960 #light ...
 </ABSTRACT-DE>
</DOC>
```

Tabelle 3.1: Übersicht über die wichtigsten Feldbelegungen der GIRT4-Kollektion. Man beachte, dass manche Felder wie Autor oder Schlagwörter mehrmals pro Dokument auftreten können, daher ist die Anzahl dort höher als die Anzahl der Dokumente.

Feldname	Anzahl der Felder pro Korpus	Anzahl der Felder pro Dokument in GIRT4-DE	Anzahl der Felder pro Dokument in GIRT4-EN
Nummer (DE=EN)	151.319	1	1
Autor (DE=EN)	237.301	1,75	1,75
Titel (DE+EN)	302.638	1	1
Deskriptoren (DE=EN)	1.535.709	10.15	10.15
Klassifikationstexte (DE=EN)	305.504	2,02	2,02
Methoden-deskriptoren	647.355	2,35	1,93
Abstracts	167.999	0,96	0,15
Schlagwörter	38.505	0,25	-
Methodentext	10.258	0,07	-

GIRT-Testanfragen

Als Testanfragen zur Evaluation des hier erstellten Retrieval-Szenarios kommen die deutschen und englischen Testanfragen zum Einsatz, die 2006 bereits für die domänenspezifische Aufgabe in CLEF verwendet wurden [Stempfhuber & Baerisch2006].

Das vorrangige Ziel bei der Erstellung dieser Testanfragen war es, ein möglichst breites Spektrum der Sozialwissenschaften abzudecken. Eine Reihe von Domain-Experten, die mit der Struktur und dem Inhalt des GIRT-Korpus vertraut waren, wurden mit der Erstellung der Anfragen betraut und reichten insgesamt 42 Vorschläge ein, von denen 25 ausgewählt wurden. Anfragen, die denjenigen der letzten Jahre zu ähnlich waren, wurden ersetzt durch Anfragen aus dem Pool der in den letzten Jahren in CLEF nicht verwendeten Vorschläge.

Die Testanfragen enthalten jeweils einen Titel, eine kurze Beschreibung über die gewünschten Informationen sowie eine ausführlichere Darstellung, die die gewünschten Informationen näher charakterisieren (siehe Listing 3.7). In diesen IR-Experimenten wurde lediglich der Titel als Anfrage verwendet. Die durchschnittliche

Anzahl der Wörter pro Titel beträgt im Englischen (Deutschen) 3,86 (3,28) Wörter.

Listing 3.7: Typische Testanfrage aus der domänenspezifischen Aufgabe in CLEF 2006

```
<num>152</num>
<DE-title>Beschäftigungspolitik auf europäischer Ebene</DE-title>
<DE-desc>Gesucht sind Informationen zur EU-Beschäftigungspolitik</DE-desc>
<DE-narr>Alle Ansätze im Bereich der Beschäftigungspolitik sind von Interesse,
        die im Zusammenhang mit der europäischen Integration stehen, sowohl auf der
        supranationalen als auch auf der nationalen Ebene
</DE-narr>
</top>
```

3.2.4 Auswahl der relevanten Datenfelder in den Testkollektionen

Nicht alle in den Dokumenten enthaltenen Informationen sind für das Information Retrieval nützlich. Beispielsweise hilft es wenig, die Autoren in den Suchindex mit aufzunehmen, wenn man sich sicher sein kann, dass nach Autoren nicht gesucht wird. Ebenso ist es - jedenfalls in diesem Testszenario - nicht notwendig, das Erscheinungsjahr eines Artikels in den Index mit aufzunehmen. Alle Testkollektionen wurden daher eingehend auf relevante XML-Felder untersucht, schließlich wurden die in Tabelle 3.2 angegebenen Felder in den Lucene-Index aufgenommen und damit der Suche verfügbar gemacht.

3.3 Lucene-Suchmaschine

Als Suchmaschine dieses IR-Szenarios wird die Suchmaschine Lucene der Apache Software Foundation[12] verwendet. Lucene ist eine Volltext-Suchmaschine, die vollständig in Java entwickelt ist und in vielen anderen Programmiersprachen ebenfalls zum freien Download zur Verfügung steht. Das Kernstück ist eine Java-Bibliothek zum Erzeugen und Durchsuchen von Indizes. Mit Hilfe dieser plattformunabhängigen Bibliothek lassen sich Freitextsuchen für beliebige Inhalte erzeugen. Eine Kerneigenschaft von Lucene ist die Möglichkeit einer hochperformanten, skalierbaren Indexierung (mehrere hundert Megabyte Daten pro Stunde). Neben der Index-Neuerstellung ist auch die Erweiterung bestehender Indices bei gleicher Indexierungsgeschwindigkeit möglich. Die Indexgröße beträgt ungefähr 30% der ursprünglichen Datenmenge.

[12] `http://lucene.apache.org/java/docs/index.html`, eingesehen im Okt. 2006

Tabelle 3.2: Übersicht über die in den Lucene-Index aufgenommenen Felder.

Kollektion	In Lucene aufgenommene Felder
OHSUMED	`Titel, Abstract, MeSH-Schlagwörter`
ImageCLEF	
- PathoPic	`Diagnose, Synonyme, Beschreibung, Klinik, Zusatzbefund, Kommentar, Info`
- Mir	`Case`
- CAS	`Description, Diagnosis, CLinicalPresentation, Keywords, Anatomy, Chapter, Title`
- Peir	`Filename, Titel, Description, RadiographType DiseaseProcess, ClinicalHistory`
GIRT	`Title, Abstract, ControlledTerm, ClassificationText`

Lucene bietet eine Vielzahl an Funktionen und Suchmöglichkeiten. Die Suche ist relevanzbasiert, das heißt die Ergebnisse werden nach Relevanz geordnet zurückgegeben. Auch andere Sortierkriterien wie datumsbasierte Sortierung sind möglich. Unterstützt wird die Suche mit booleschen Operatoren, ungenaue Suche, Suche mit Platzhaltern (Wildcards), Distanzsuche, Feldsuche oder auch Suche mit Verstärkungsfaktoren. Die wichtigsten Suchfunktionen werden nun kurz beschrieben:

- **Boolesche Operatoren**: Eine Suchanfrage in Lucene ist in Ausdrücke und Operatoren unterteilt. Ausdrücke sind einzelne Suchwörter (Bsp. "Überblick") oder Phrasen ("Überblick Sozialwissenschaften"). Mehrere Ausdrücke werden mit Booleschen Operatoren miteinander verknüpft. Zu den Booleschen Operatoren zählen `OR, AND` und `NOT`. `OR` ist die Standardverknüpfung in Lucene und kann auch weggelassen werden (Bsp. "foo bar" entspricht "foo `OR` bar"). `AND` kann auch durch ein Pluszeichen ersetzt werden (Bsp. "+foo +bar")). `NOT` entspricht dem Minuszeichen und muss immer mit nicht negierten Ausdrücken kombiniert werden (Bsp. nur "-foo" ist unzulässig). Durch Angabe von Klammern ist eine Gruppierung der Suchanfrage möglich (Bsp. "-foo +(bar foobar)" sucht nach "bar" oder "foobar" und schließt "foo" gleichzeitig aus).

- **Ungenaue Suche**: Durch die Implementierung des Levenshtein Distanz Algorithmus ([Levenshtein1966]) wird eine ungenaue Suchanfrage (Fuzzy-Suche) in Lucene ermöglicht. In der Anfrage wird hierzu wird eine Tilde verwendet (Bsp. "suchen∼"). Das Dokument sucht dann nach ähnlichen Wörtern (Bsp.

"buchen" oder suche"). Ausdrücke, die über ungenaue Suche gefunden werden, werden automatisch mit dem Faktor 0,2 gewichtet.

- **Suche mit Platzhaltern**: Lucene erlaubt die Wildcard-Suche für ein oder mehrere Zeichen. Ein Platzhalter für mehrere Zeichen wird mit dem Stern-Symbol angegeben (Bsp. "bau*" findet "bau, baum, bauch, ..."), ein Platzhalter für ein einzelnes Zeichen mit einem Fragezeichen (Bsp. "a?t" findet "abt, ast, axt, ...").

- **Distanzsuche**: Für mehrere Wörter, die in der Suchanfrage vorkommen, kann eine bestimmte maximale Entfernung dieser Wörter zueinander angegeben werden. Hierzu werden die Wörter in Hochkommata gesetzt, gefolgt von einer Tilde, hinter der die maximale Distanz zwischen diesen Wörtern angegeben ist (Bsp. "Professor Mustermann"$\sim$ 3 findet "Professor Mustermann" und " Professor Dr. Mustermann").

- **Feldsuche**: Lucene unterstützt die Erstellung von Indexfeldern wie *Dokumententitel* und *Inhalt*. Bei der Suchanfrage können dementsprechend Feldnamen angegeben werden, auf die die Suche beschränkt ist. Bei der Suche nach "title:(foo bar)" (äquivalent zu "title:foo title:bar") muss entweder "foo" oder "bar" im Titel gefunden werden.

- **Suche mit Verstärkungsfaktoren**: In Lucene können Ausdrücke (Wörter und Phrasen) mit Verstärkungsfaktoren versehen werden, was sich auf die relevanzbasierte Sortierung der Suchergebnisse auswirkt. Zur Verwendung eines Verstärkungsfaktors wird das "$\wedge$"-Symbol, gefolgt von einem Zahlenwert hinter dem Ausdruck angegeben. Der Standardfaktor ist 1. In dem Beispiel "foo$\wedge$4 bar" wird "foo" viermal so stark gewichtet wie "bar"

Der Gewichtungsalgorithmus von Lucene basiert auf dem Produkt aus Häufigkeit $TF(t,d)$ eines Terms t innerhalb eines Dokumentes d und der inversen Häufigkeit des Terms in der gesamten Dokumentenkollektion ($IDF(t)$). Zusätzlich berücksichtigt werden bei der Gewichtung der Verstärkungsfaktor $boost(t,d)$ einzelner Ausdrücke, ein Normalisierungsfaktor $norm(d)$ der Dokumente, der Anteil $frac(t,d)$ eines Ausdrucks im Dokument sowie ein Normalisierungsfaktor $norm(q)$ der Anfrageterme. Das Gewichtungsergebnis einer Anfrage wird damit wie folgt berechnet:

$$score(q,d) = \sum_{tinq} tf(t,d) * idf(t) * boost(t,d) * norm(t,d) * frac(t,d) * norm(q). \quad (3.1)$$

3.4 Experimentelles Szenario

Bei der Implementierung eines Standard-Systems zur Dokumentenrecherche können vereinfachend zwei Phasen unterschieden werden. In der *Indexierungsphase* werden die Dokumentenkollektionen eingelesen und die notwendigen Textpassagen aus den Dokumenten extrahiert. Die Texte werden dann auf die gewünschten Arten vorverarbeitet (beispielsweise können Stoppwörter entfernt werden) und schließlich in einen Suchmaschinenindex geschrieben. In der *Suchphase* können Anfragen eines Benutzers oder einer Anwendung an die Suchmaschine gestellt werden. Die Anfrage wird auf die gleiche Weise wie die Dokumente vorverarbeitet und in die Anfragesprache der Suchmaschine überführt. Bei der anschließenden Abfrage ermöglicht der in der Indexierungsphase erstellte Index eine schnelle Ermittlung der Ergebnisse.

In dieser Arbeit sind alle Applikationen, die für das IR-Szenario entwickelt wurden, vollständig in die sogenannte *Unstructured Information Management Architecture* (UIMA) eingebettet. UIMA ist ein Rahmenwerk für rechnergestützte Anwendungen zur natürlichen Sprachverarbeitung, welches eine einfache Integration und Kombination verschiedenster NLP-Anwendungen über eine gemeinsame Schnittstelle ermöglicht. *UIMA* wird im folgenden Abschnitt kurz erläutert, bevor im Anschluss die verschiedenen Testszenarien näher beschrieben werden.

3.4.1 *Unstructured Information Management Architecture*

Die enorme Fülle wissenschaftlicher Informationen in Form von Texten, Filmen oder Sprachaufzeichnungen stellt nicht nur für das Information Retrieval, sondern auch für andere Bereiche der natürlichen Sprachverarbeitung wie Text-Mining eine besondere Herausforderung dar. Wissenschaftler und Unternehmen stellen dieser Herausforderung eine Vielzahl heterogener Anwendungen entgegen, die diese unstrukturierten Informationen analysieren, organisieren und relevantes Wissen an den Benutzer zurückliefern (siehe Abbildung 3.3). Zu diesen Techniken zählen statistische oder regelbasierte Verfahren zur natürlichen Sprachverarbeitung, maschinelles Lernen, automatisches Schlussfolgern (engl. Automated Reasoning) oder auch maschinelles Übersetzen. In der Regel konzentrieren sich diese sogenannten *Unstructured Information Management (UIM) Applikationen* [Ferrucci & Lally2004a] auf einzelne Teilaspekte der Sprachverarbeitung. Auch die in diesem IR-Szenario verwendeten Porter-Stemmer sowie das MorphoSaurus-System können als UIM-Applikationen angesehen werden, da sie jeweils eine spezifische NLP-Aufgabe erfüllen.

Die *Unstructured Information Management Architecture* (UIMA) wurde von IBM Research entwickelt und stellt eine Architektur und ein Rahmenwerk bereit, welche

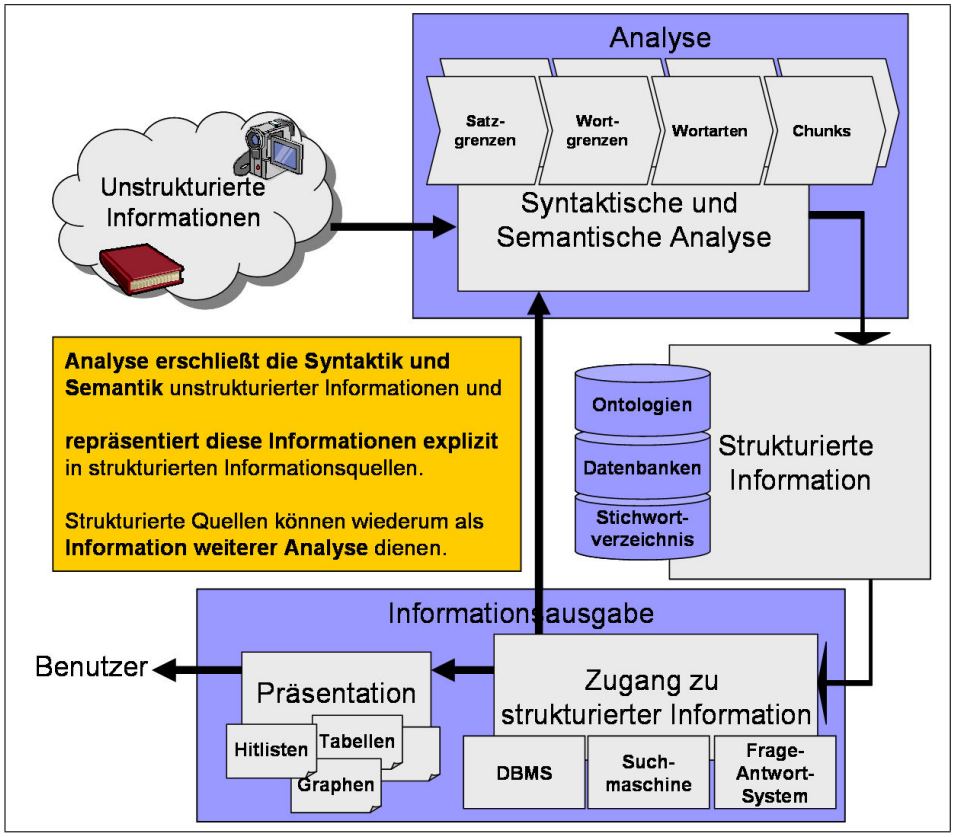

Abbildung 3.3: Der Weg von unstrukturierter zur strukturierten Information. Eine typische UIM-Applikation aus dem Bereich der natürlichen Sprachverarbeitung (DBMS: Datenbankmanagementsystem) (übersetzt und modifiziert aus [UIMA2006]).

die Integration verschiedenster UIM-Anwendungen ermöglichen und so die unstrukturierte und strukturierte Informationswelt miteinander verknüpfen (siehe Abbildung 3.4). Entwickler von UIM-Anwendungen können ihre NLP-Anwendungen in das UIMA-Rahmenwerk einbetten und diese somit über gemeinsame Schnittstellen austauschen oder miteinander verbinden. Die UIMA-*Architektur* beschreibt die Schnittstellen, Datenrepräsentationen, Entwurfsmuster und Entwicklungsregeln zur Entwicklung, Verknüpfung und zum Anwenden von UIM-Anwendungen. Das *UIMA-Rahmenwerk* stellt eine plattformunabhängige Laufzeitumgebung zur Integration und Komposition verschiedener Komponenten bereit. Das *UIMA Software Development Kit (SDK)* enthält eine vollständig in Java basierte Implementierung des Rahmenwerks. Für eine ausführliche Beschreibung über die Verwendung von UIMA wird auf [Ferrucci & Lally2004b, UIMA2006] sowie auf die UIMA JavaDocs verwiesen. In diesem Abschnitt können lediglich die zentralen UIMA-Konzepte kurz angesprochen werden, die auch in diesem IR-Szenario Anwendung finden.

Die von Entwicklern entworfenen NLP-Anwendungen werden in UIMA *Anno-*

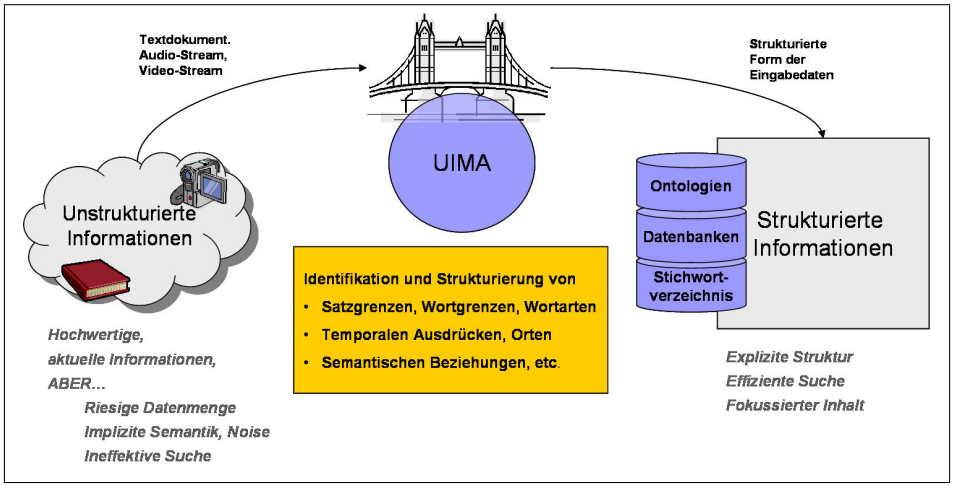

Abbildung 3.4: UIMA bildet die Brücke zwischen unstrukturierter und strukturierter Information (übersetzt und modifiziert aus [UIMA2006]).

tatoren genannt. *Porter-Stemmer* und MorphoSaurus sind solche Annotatoren. Durch die Einbettung der Annotatoren in das UIMA-Rahmenwerk werden aus den Annotatoren sogenannte *Analysis Engines*. *Analysis Engines* erweitern die Annotatoren um die notwendigen Schnittstellen, so dass die Annotatoren im UIMA Rahmenwerk ausgeführt werden können. Darüber hinaus erzeugen *Analysis Engines* Metainformationen in Form von XML-Dateien, die das Dokument oder bestimmte Textbereiche näher beschreiben. Metainformationen sind alle im Verlauf der Analyse erzeugten Informationen. Werden beispielsweise in einem ersten Schritt der Analyse XML-Dokumente ausgelesen und die darin enthaltenen Informationen wie Dokumenten-ID, Namen der Autoren oder Erstellungsjahr extrahiert, so sind dies bereits die ersten Metainformationen. Wird anschließend der enthaltene Text an ein Analysewerkzeug wie den Porter-Stemmer weitergereicht, fügt diese Anwendung dem Dokument weitere Metainformation hinzu, im Falle des Porter-Stemmers die in Stammform überführte Form des Dokumentes.

Ein Ziel von UIMA ist es, mit möglichst minimalem Aufwand verschiedene *Analysis Engines* über eine einheitliche Schnittstelle zu sogenannten *Aggregierten Analysis Engines* zu kombinieren. Hierfür ist keine Implementierarbeit erforderlich, es wird lediglich ein XML-Dokument erstellt, in dem die Namen und die Reihenfolge der einzelnen *Analysis Engines* angegeben wird. In dem vorgestellten IR-Szenario werden beispielsweise ein XML-Parser, der die Informationen aus den XML-Dateien ausliest, ein Stoppwort-Entferner, der Porter-Stemmer sowie das MorphoSaurus-System zu einer *Aggregierten Analysis Engine* kombiniert. Zusätzlich zu der *Aggregierten Analysis Engine* werden noch zwei Anwendungen benötigt, die die Textdateien aus dem Dateisystem auslesen und nach der Analyse in strukturierter Form,

beispielsweise als Suchmaschinenindex, in das Dateisystem zurück schreiben. Diese zwei Komponenten sind der *Collection Reader*, der die Dateien einliest, und *CAS Consumer*[13], welcher die Dokumente wieder in das Dateisystem zurück schreibt. Die Komponenten *Collection Reader, (Aggregierte) Analysis Engine* und *CAS Consumer* bilden zusammen die sogenannte *Collection Processing Engine (CPE)*. Die CPE stellt also eine Art Pipeline dar, in die unstrukturierte Dokumente eingelesen werden, die anschließend von einem oder mehreren *Analysis Engines* analysiert werden und in strukturierter Form wieder im Dateisystem gespeichert werden (Abbildung 3.5).

3.4.2 Einbettung des IR-Szenarios in UIMA

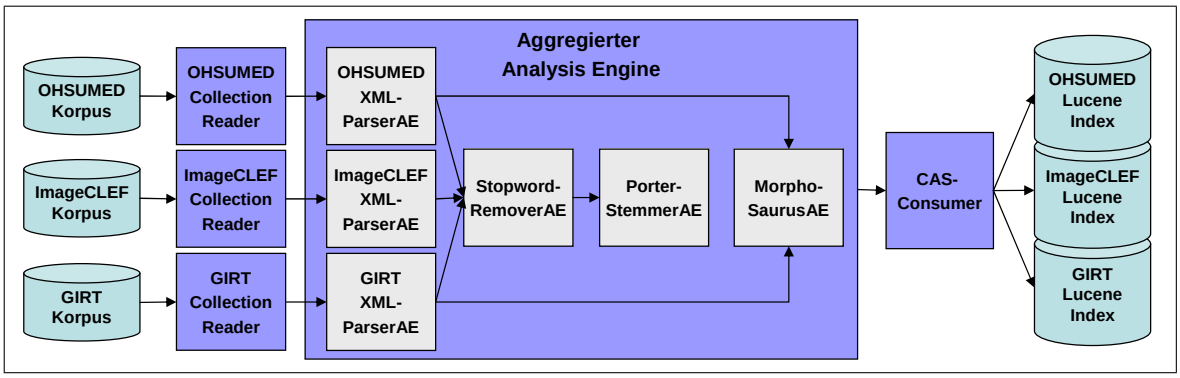

Abbildung 3.5: Die Collection Processing Engine für das IR-Szenario. Für jede der drei Kollektionen stehen eigene *Collection Reader* zur Verfügung. Diese übergeben das eingelesene Dokument der *Aggegrierten Analysis Engine*, die aus vier einzelnen *Analysis Engines* besteht. Diese werden der Reihe nach durchlaufen. Anschließend werden die Dokumente einem allgemeinen *CAS-Consumer* übergeben, der die Daten in einen Lucene-Index schreibt.

Die Indexierungsphase des IR-Szenarios ist in Form einer *Collection Processing Engine* realisiert (siehe Abbildung 3.5). Die in der Abbildung dargestellte Reihenfolge entspricht der Reihenfolge, in der ein Dokument die CPE durchwandert. Am Beginn der CPE-Pipeline stehen die dokumentenspezifischen *Collection Reader*, welche die drei Dokumentenkollektionen OHSUMED, ImageCLEF und GIRT aus dem Dateisystem einlesen. Da die Kollektionen in unterschiedlichem Datenformat vorliegen, muss für jede Kollektion ein eigener Reader entwickelt werden. Die Reader

[13] *CAS* steht für *Common Analysis Structure*. Hiermit wird im Wesentlichen das analysierte Artefakt samt Metainformationen bezeichnet, welches durch den *Consumer* in strukturierter Form in das Dateisystem zurück geschrieben wird.

übergeben jedes Dokument aus der Dokumentenkollektion einzeln der *Aggregierten Analysis Engine*. Diese Engine ist aus vier Teilkomponenten aufgebaut:

- **XML-Parser**: Zunächst wird das Dokument einem XML-Parser übergeben, der das XML-basierte Dokument in Freitext überführt und bereits die ersten Metainformationen erzeugt. Je nach verfügbaren XML-Tags (siehe Abschnitt 3.2) können dies zum Beispiel Autorenname, Erstellungsjahr oder Dokumentensprache sein. Sämtliche relevante Freitextinformationen werden anschließend den weiteren Analyseanwendungen übergeben.

- **Stoppwort-Entferner**: Wie bereits beschrieben, soll in diesem IR-Szenario die klassische Freitextsuche mit einem konventionellen Stemming-Verfahren (Porter-Stemmer) sowie mit dem MorphoSaurus-System verglichen werden. Bei der klassischen Freitextsuche werden lediglich die Stoppwörter entfernt. Stoppwörter nennt man im Information Retrieval diejenigen Wörter, die bei einer Volltextindexierung nicht beachtet werden, da sie sehr häufig auftreten und gewöhnlich keine Relevanz für die Erfassung des Dokumentinhalts besitzen (wie "und" und "oder"). Die in diesem Szenario verwendeten Stoppwörter entstammen der Snowball-Implementierung des Porter-Stemmers und beinhalten 174 englische, 234 deutsche und 155 französische Stoppwörter[14].

- **Porter-Stemmer**: Die von den Stoppwörtern befreite Version des Dokumentes wird anschließend dem Porter-Stemmer übergeben. In diesem Szenario wird der Snowball-Stemmer[15], der sprachspezifische Implementierungen des Porter-Algorithmus bereitstellt.

- **MorphoSaurus**: Das MorphoSaurus-System ist die vierte *Analyis Engine* in der Pipeline. Es erhält die Dokumente direkt von dem XML-Parser und umgeht somit den Stoppwort-Entferner, da es seine eigene Stoppwortliste enthält. Das MorphoSaurus-System wandelt den Freitext in die sprachunabhängige, morpho-semantisch normalisierte Form um.

Nach Abschluss der Analyse liegen die Dokumente schließlich in fünf unterschiedlichen Versionen vor, und zwar in XML-Format, in Freitext, in Freitext abzüglich der Stoppwörter, in Stammform sowie in morpho-semantisch normalisierter Form. Diese Versionen werden dem letzten Glied in der Pipeline weitergereicht, dem *CAS-Consumer*, welcher den Lucene-Index erzeugt. Für jede der drei Kollektionen OHSU-

[14] `http://snowball.tartarus.org/algorithms/<english,german,french>/stop.txt`, eingesehen im Februar 2007

[15] `http://snowball.tartarus.org/` verwendet, eingesehen im Februar 2007

MED, ImageCLEF und GIRT werden eigene Lucene-Indizes erzeugt. In den Lucene-Indices werden für jedes Dokument folgende Felder erstellt:

- `docid`: enthält die eindeutige Identifikationsnummer des Dokumentes.
- `lang`: enthält die Sprache, in der die Dokumente verfasst sind.
- `orig`: enthält sämtliche im Dokument relevanten Textpassagen in der von der Stoppwort befreiten Form. Dabei werden die verschiedenen Teile des Dokumentes wie Titel, Abstract, Schlagwörter, Beschreibung etc. durch einfaches Leerzeichen getrennt aneinandergefügt.
- `stem`: enthält sämtliche relevanten Textpassagen in ihrer Stammform.
- `mid`: enthält die relevanten Textpassagen in ihrer morpho-semantisch normalisierten Form.

Anmerkend sei gesagt, dass die einzelnen XML-Tags aus den Dokumenten wie Titel, Abstract oder Schlagwörter prinzipiell auch als separate Felder im Lucene-Index definiert werden können, die dann von Lucene einzeln und mit unterschiedlicher Gewichtung durchsuchbar sind. Die Gesamtergebnisse aller im Folgenden beschriebenen Testläufe würden dadurch vermutlich sogar gesteigert. Das vorrangige Ziel dieser Arbeit ist jedoch der Vergleich der verschiedenen Verfahren zur Sprachverarbeitung unter standardisierten Bedingungen, weniger die Optimierung der Lucene-Einstellungen auf diese Dokumentenkollektionen. Daher wurde in den Experimenten auf eine kompliziertere Indexstruktur verzichtet.

3.4.3 Experimentelle Testläufe

Nachdem in der Indexierungphase verschiedene Lucene-Indices für die Dokumente erstellt wurden, können nun Testanfragen an die Indices gerichtet werden. Alle Anfragen aus den Testkollektionen werden von derselben *Aggegierten Analysis Engine*, die auch in der Indexierungsphase verwendet wurde, in die Darstellungen *Orig, Stem* und *Mid* überführt. Wie bereits beschrieben ist es das vorrangige Ziel der Testläufe herauszufinden, ob und in welchem Umfang morphologische Analyse die Retrieval-Performanz verbessert. Daher werden die verschiedenen Darstellungen der Testanfragen getrennt und in Kombination an den Lucene-Index geschickt und die Retrieval-Performanz gemessen. Die Testläufe werden so benannt, dass an erster Stelle die jeweilige Version der Anfrage steht (*Orig, Stem* oder *Mid*), gefolgt von der Sprache der Testanfrage (*En* oder *De*) sowie der Sprache der Dokumente (*En* oder

De). Tabelle 3.3 gibt einen Überblick über die verschiedenen Sprachen, in denen für die verschiedenen Kollektionen Testläufe durchgeführt wurden. Im Folgenden werden die Testläufe für das Deutsche näher vorgestellt, die englischen und französischen Testanfragen sind entsprechend aufgebaut:

1. *Orig-De-De*: Als Baseline wird die deutsche Original-Anfrage verwendet, aus denen lediglich die Stoppwörter entfernt wurden. Gesucht wird im Lucene-Feld *orig*.

2. *Stem-De-De*: Die Testanfragen, welche vom Porter-Stemmer verarbeitet wurden, werden als zweiter Testlauf an die entsprechenden *stem*-Felder im Lucene-Index adressiert.

3. *Mid-De-De*: In diesem Testlauf werden die *MID*-Anfragen der MorphoSaurus-Repräsentation verwendet. Mit diesen Anfragen wird in den *mid*-Feldern des Lucene-Index gesucht.

4. *OrigStem-De-De*: In diesem Testlauf werden die Originalanfrage und die gestemmte Anfrage kombiniert (einfache *ODER*-Verknüpfung) und auf den jeweiligen Feldern im Lucene-Index gesucht.

5. *OrigMid-De-De*: Hier wird die Originalanfrage mit der MorphoSaurus-Anfrage kombiniert. Die Felder, auf denen im Lucene-Index gesucht wird, sind *orig* und *mid*.

6. *StemMid-De-De*: In diesem Run wird die gestemmte Anfrage mit der MorphoSaurus-Anfrage kombiniert. Die Felder, auf denen im Lucene-Index gesucht wird, sind demnach *stem* und *mid*.

7. *OrigStemMid-De-De*: Dieser Testlauf berücksichtigt schließlich alle drei Anfrage-Versionen und sucht dementsprechend über allen drei Feldern im Lucene-Index.

Da die OHSUMED-Kollektion nur in englischer Sprache vorliegt, wurden auf dieser Kollektion Testläufe in englischer Sprache durchgeführt. Für die ImageCLEF-Kollektion wurden der trilingualen Dokumentkollektion entsprechend Testläufe in englischer, deutscher und französischer Sprache implementiert. Die Testläufe für die deutsch-englische GIRT-Kollektion wurden in deutscher und englischer Sprache durchgeführt. Für die GIRT-Dokumentenkollektion sind typische Beispielanfragen für jeden dieser Testläufe in Beispiel-Box 1 angegeben.

Tabelle 3.3: Übersicht über die Sprachen, in denen Testläufe für die verschiedenen Kollektionen durchgeführt wurden.

Kollektion	Sprache der Testläufe
OHSUMED	Englisch
ImageCLEF	Englisch, Deutsch, Französisch
GIRT	Englisch, Deutsch

Beispiel Box 1: Anfragerepräsentation für verschiedene Testläufe aus den GIRT-Testanfragen. Disjunktionen (ODER) müssen in der Lucene-Anfrage nicht explizit markiert werden, Konjunktionen (AND) können durch ein + Symbol ausgedrückt werden.

Anfrage: "Beschäftigungspolitik auf deutscher Ebene"

Beispiel 1 – Run: Orig-De-De (Baseline)
(+lang:de +orig:beschäftigungspolitik) (+lang:de +orig:deutscher) (+lang:de +orig:ebene)

Beispiel 2 – Run: Stem-De-De
(+lang:de +stem:beschaeftigungspolitik) (+lang:de +stem:deutsch) (+lang:de +stem:eben)

Beispiel 3 – Run: Mids-De-De
(+lang:de +mid:job) (+lang:de +mid:policy) (+lang:de +mid:german) (+lang:de +mid:level)

Beispiel 4 – Run: OrigStem-De-De
(+lang:de +orig:beschäftigungspolitik) (+lang:de + orig:deutscher) (+lang:de + orig:ebene) (+lang:de +stem:beschaeftigungspolitik) (+lang:de +stem:deutsch) (+lang:de +stem:eben)

Beispiel 5 – Run: OrigMid-De-De
(+lang:de +orig:beschäftigungspolitik) (+lang:de + orig:deutscher) (+lang:de + orig:ebene) (+lang:de +mid:job) (+lang:de +mid:policy) (+lang:de +mid:german) (+lang:de +mid:level)

Beispiel 6 – Run: StemMid-De-De
(+lang:de +stem:beschaeftigungspolitik) (+lang:de +stem:deutsch) (+lang:de +stem:eben) (+lang:de +mid:job) (+lang:de +mid:policy) (+lang:de +mid:german) (+lang:de +mid:level)

Beispiel 7 – Run: OrigStemMid-De-De
(+lang:de +orig:beschäftigungspolitik) (+lang:de + orig:deutscher) (+lang:de + orig:ebene) (+lang:de + stem:beschaeftigungspolitik) (+lang:de + stem:deutsch) (+lang:de + stem:eben) (+lang:de +mid:job) (+lang:de +mid:policy) (+lang:de +mid:german) (+lang:de +mid:level)

3.5 Ergebnisse des IR-Szenarios

3.5.1 Bewertung der Effektivität eines IR-Systems

Bei der Evaluation von IR-Systemen wird bestimmt, wie viele relevante Dokumente von dem System gefunden werden und möglichst weit vorne in der Trefferliste aufgelistet werden. Dabei sollen die Anfragen präzise und zugleich erschöpfend beantwortet werden. Das zentrale Problem hierbei ist, dass die richtige Antwort bekannt sein muss, um die Antwort des Systems bewerten zu können. In der Regel werden in der Praxis Experten gebeten, die Relevanz von Dokumenten bezüglich einer Anfrage einzuschätzen. Aufbauend auf diesen Urteilen über die Relevanz von Dokumenten werden die zwei am häufigsten verwendeten Evaluierungsmaße definiert, die Aussagen über die Präzision und die Vollständigkeit eines Suchergebnisses und somit über die Güte des IR-Systems treffen. Vollständigkeit und Präzision werden im Information Retrieval als *Recall* und *Precision* bezeichnet. In Formeln ausgedrückt ist die Precision P definiert als

$$P = \frac{Zahl\ der\ gefundenen\ relevanten\ Dokumente}{Zahl\ aller\ gefundenen\ Dokumente}$$

und der Recall R als

$$R = \frac{Zahl\ der\ gefundenen\ relevanten\ Dokumente}{Zahl\ aller\ relevanten\ Dokumente}$$

Im Zuge der Evaluationskampagnen wie TREC und CLEF wurde die Forderung nach einem einzigen Gütewert laut, welcher die objektive Vergleichbarkeit von Systemen ermöglichen sollte. So entwickelte sich die *Mean Average Precision (MAP)* zum bekanntesten Wert, um verschiedene Systeme anhand eines Wertes miteinander zu vergleichen. Er stellt das Mittel aus den Durchschnitts-Präzisionen MAP_i aller Testanfragen i dar. MAP_i wiederum ist der Durchschnitt aller Präzisionswerte einer Testanfrage an genau denjenigen Cut-Off-Punkten, an denen ein relevantes Dokument gefunden wurde. Die durchschnittliche Präzision MAP_i einer Anfrage i errechnet sich nach folgender Formel:

$$\mathrm{MAP_i} = \frac{\sum_{r=1}^{N_{retr}} (P(r) \times \mathrm{rel}(r))}{N_{rel}}, \quad mit$$

r - Rang des Dokumentes in der Ergebnisliste
$\mathrm{rel}(r)$ - Binärfunktion der Relevanz an Stelle r
N_{retr} - Anzahl der gefundenen Dokumente
$P(r)$ - Präzisionswert an der Stelle r
N_{rel} - Anzahl der relevanten Dokumente

P5 und P20 bezeichnen die Präzisionswerte an den Cut-Off-Punkten 5 und 20. Ein Cut-Off-Wert gibt die Anzahl der gefundenen Dokumente an, die bei der Berechnung des Präzisionswertes berücksichtigt werden. P5 und P20 stellen nach Meinung des Autors ein wichtiges Maß für die Qualität von Recherchesystemen dar, da ein Benutzer eines Suchdienstes meist nicht mehr als die ersten 20 Ergebnisse einer Ergebnisliste betrachtet, bevor er die Suche beendet oder die Anfrage ändert. In den folgenden Ergebnissen der Testläufe sind dementsprechend der MAP-Wert sowie die Werte P5 und P20 angegeben. Zur Signifikanzüberprüfung wurde der Vorzeichen-Rangtest nach Wilcoxon durchgeführt.

3.5.2 OHSUMED - Ergebnisse

Die Ergebnisse der Testläufe auf der OHSUMED-Kollektion sind in Tabelle 3.4 angegeben. Die Präzisionswerte an verschiedenen Cut-Off-Punkten sowie der MAP-Wert sind außerdem in der Abbildung 3.6 grafisch dargestellt. In dieser Testkollektion schneidet das MorphoSaurus-System bezüglich des MAP-Wertes mit 0,1745 (8% besser als die Baseline) von den singulären Verfahren am besten ab, allerdings ist diese Erhöhung nicht signifikant (Wilcoxon, $\alpha = 0,05$). Die kombinierten Verfahren sind deutlich besser als die einzelnen Verfahren, der MAP-Wert des *OrigStemMid*-Testlaufs ist mit 0,1991 (+23% gegenüber der Baseline) am höchsten.

Auch bezüglich der P5 und P20 Werte sind die drei Verfahren relativ dicht beieinander. *Orig* liegt bei P5 mit 0,4305 etwas vor *Stem* mit 0,4133, gefolgt von *Mid* mit 0,3943. Dies bedeutet, dass *Orig* unter den ersten fünf Ergebnissen mehr relevante Treffer findet als die beiden anderen Verfahren. Für P20 ist die Reihenfolge wiederum umgekehrt, hier liegt *Mid* mit 0.3024 leicht vor *Stem* und *Orig* mit 0,2867 und 0,2838. Auch bei P5 und P20 schneiden die kombinierten Verfahren besser ab als alle singulären Verfahren, die beiden besten Testläufe sind *OrigMid* und *OrigStemMid* mit P5 von jeweils 0,4667 und P20 von 0,33 bzw. von 0,3286.

3.5.3 ImageCLEF - Ergebnisse

Eine Übersicht über die Ergebnisse der ImageCLEF-Dokumentenkollektion gibt Tabelle 3.5. Wiederum sind verschiedene Präzisionswerte sowie der MAP-Wert für die englischen Ergebnisse grafisch dargestellt (siehe Abbildung 3.7). Zunächst fällt auf, dass die MAP-Werte für die deutschen und die französischen Testläufe extrem niedrig sind. Dies hängt damit zusammen, dass die Anzahl der deutschen und französischen Dokumente mit 9.000 deutschen und 2.000 französischen Bildunterschriften im Vergleich zur englischen Kollektion mit ca. 40.000 Bildunterschriften

Tabelle 3.4: Übersicht über die Ergebnisse der OHSUMED-Testläufe. Angegeben sind jeweils der MAP sowie P5 und P20. Signifikante Veränderungen (Wilcoxon, $\alpha = 0,05$) sind mit Sternchen (*) gekennzeichnet.

Szenario	Testlauf	MAP	Abw. (in %)	p-Wert (Wilcoxon)	P5	P20
Englisch	Orig-En-En	0,1620			0,4305	0,2838
	Stem-En-En	0,1662	+2,59	0,6042	0,4133	0,2867
	Mid-En-En	0,1745	+7,72	0,3873	0,3943	0,3024
	OrigStem-En-En	0,1757	+8,46	<0,0001*	0,4381	0,3019
	OrigMid-En-En	0,1959	+20,93	<0,0001*	**0,4667**	**0,3300**
	StemMid-En-En	0,1963	+21,17	<0,0001*	0,4457	0,3224
	OrigStemMid-En-En	**0,1991**	+22,90	<0,0001*	**0,4667**	0,3286

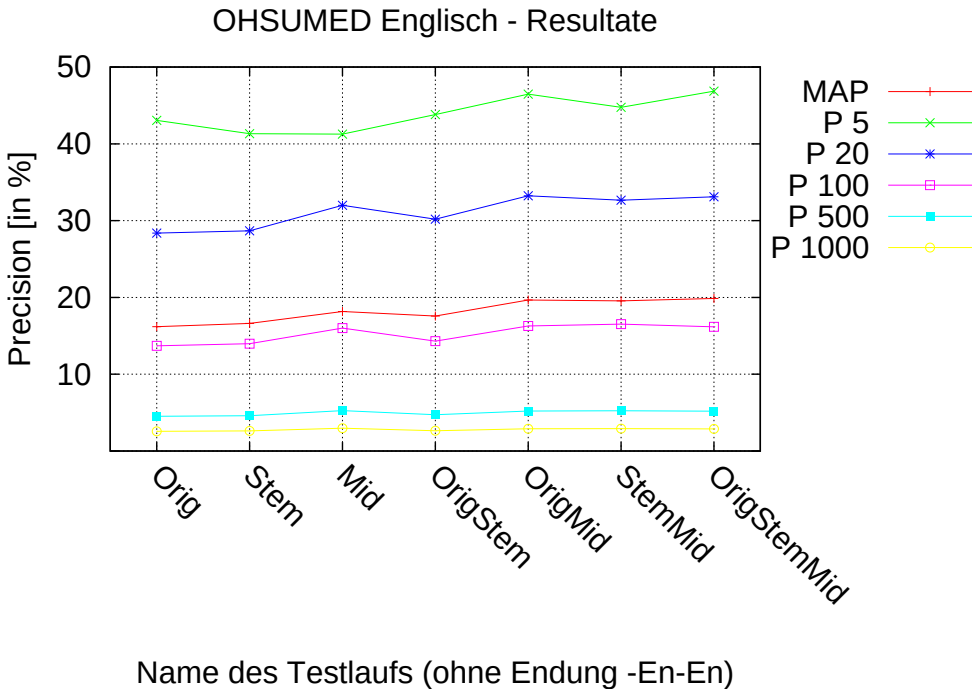

Abbildung 3.6: MAP sowie Precision-Ergebnisse der englischen OHSUMED-Testläufe bei verschiedenen Cut-Off-Punkten.

relativ klein ist. Laut mündlichen Informationen der Organisatoren von ImageCLEF wurde auch nur ein Teil der deutschen und französischen Dokumente auf Relevanz beurteilt, die restlichen wurden pauschal als nicht relevant beurteilt. Daher sind die Ergebnisse fürs Deutsche und Französische nur bedingt aussagekräftig und nicht statistisch signifikant. Tendenziell sind die MAP-Ergebnisse im Deutschen mit dem MorphoSaurus-System etwas besser als ohne das System, im Französischen hingegen scheint es weniger Nutzen zu bringen. Dies korreliert mit der Abdeckung und Qualität der MorphoSaurus-Lexika, welche für das Deutsche bereits sehr umfangreich und für das Französische noch deutlich ausbaufähig sind. Ansonsten sollen diese Ergebnisse an dieser Stelle jedoch nicht weiter interpretiert werden.

Für das Englische liegt das MorphoSaurus-System mit einem MAP-Wert von 0,1487 in dieser Kollektion um 12% unter der Baseline (allerdings nicht statistisch signifikant) und um 16,2% unter der Stammform des Porter-Stemmers. Die kombinierten Verfahren sind den einzelnen Verfahren wiederum überlegen, am besten und statistisch signifikant schneidet die Kombination aller drei Verfahren mit einem MAP-Wert von 0,1943 (+14,97% gegenüber der Baseline) ab. Bezüglich P5 ist *Stem* mit 0,5333 vor *Mid* mit 0,50 und *Orig* mit 0,4867. Die kombinierten Verfahren schneiden etwas besser ab, den höchsten P5-Wert erreicht *StemMid* mit 0,5533.

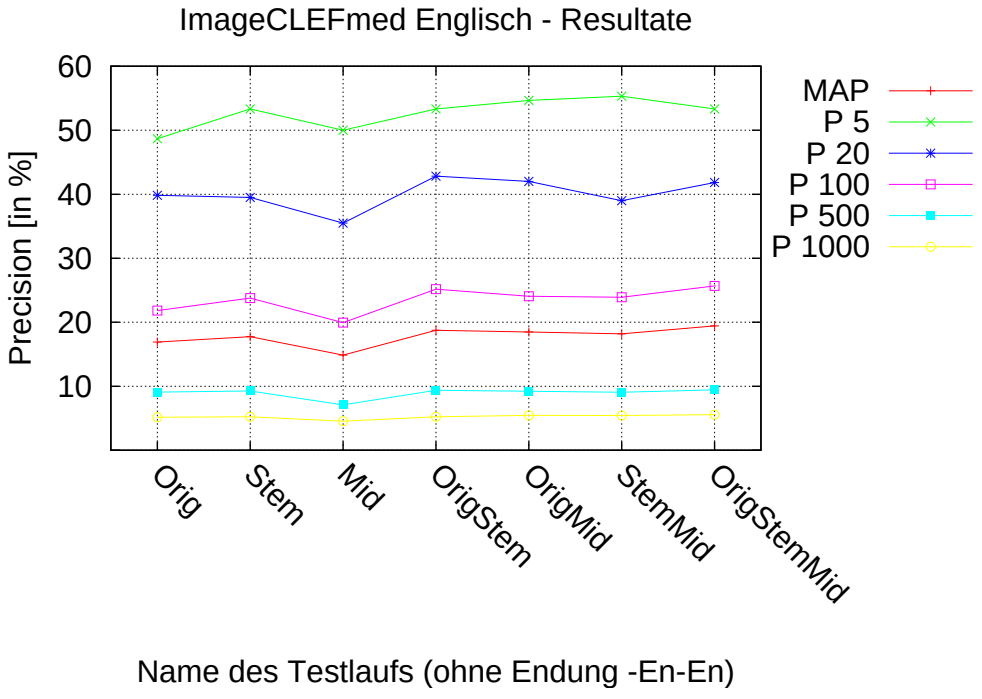

Abbildung 3.7: MAP sowie Precision-Ergebnisse der englischen ImageCLEF-Testläufe bei verschiedenen Cut-Off-Punkten.

Tabelle 3.5: Übersicht über die ImageCLEF Ergebnisse für Deutsch und Englisch. Angegeben sind jeweils der MAP sowie P5 und P20. Signifikante Veränderungen (Wilcoxon, $\alpha = 0,05$) sind mit Sternchen (*) gekennzeichnet.

Szenario	Testlauf	MAP	Abw. (in %)	p-Wert (Wilcoxon)	P5	P20
Deutsch	Orig-De-De	0,0335			0,1760	0,1500
	Stem-De-De	0,0343	+2,39	0,8445	0,1857	0,0964
	Mid-De-De	0,0403	+20,30	0,1769	0,1400	0,0750
	OrigStem-De-De	0,0376	+12,24	0,9687	**0,2071**	**0,1268**
	OrigMid-De-De	**0,0449**	+34,03	0,2762	0,1933	0,1183
	StemMid-De-De	0,0407	+21,49	0,2365	0,1600	0,0650
	OrigStemMid-De-De	0,0446	+33,13	0,3437	0,2000	0,1167
Englisch	Orig-En-En	0,1690			0,4867	0,3983
	Stem-En-En	0,1775	+5,03	0,9910	0,5333	0,3950
	Mid-En-En	0,1487	-12,01	0,3525	0,5000	0,3550
	OrigStem-En-En	0,1875	+10,95	0,0396*	0,5333	**0,4283**
	OrigMid-En-En	0,1847	+9,29	0,0798	0,5467	0,4200
	StemMid-En-En	0,1819	+7,63	0,1247	0,5533	0,3900
	OrigStemMid-En-En	**0,1943**	+14,97	0,021*	**0,5333**	0,4183
Französisch	Orig-Fr-Fr	0,0199			0,1655	0,1431
	Stem-Fr-Fr	**0,0263**	+32,16	0,1832	**0,1933**	0,1500
	Mid-Fr-Fr	0,0198	-0,50	0,7960	0,1867	0,1383
	OrigStem-Fr-Fr	0,0236	+18,59	0,1173	0,1667	**0,1517**
	OrigMid-Fr-Fr	0,0188	-5,53	0,8433	0,1867	0,1283
	StemMid-Fr-Fr	0,0211	+6,03	0,6814	0,1867	0,1383
	OrigStemMid-Fr-Fr	0,0204	+2,51	0,5531	0,1667	0,1317

Tabelle 3.6: Übersicht über die GIRT-Ergebnisse für Deutsch und Englisch. Angegeben sind jeweils der MAP sowie P5 und P20. Signifikante Veränderungen (Wilcoxon, $\alpha = 0,05$) sind mit Sternchen (*) gekennzeichnet.

Szenario	Testlauf	MAP	Abw. (in %)	p-Wert (Wilcoxon)	P5	P20
Deutsch	Orig-De-De	0,2777			0,6320	0,5400
	Stem-De-De	0,3535	+27,30	0,0018*	0,7040	0,6280
	Mid-De-De	0,2846	+2,48	0,6528	0,5200	0,4900
	OrigStem-De-De	0,3530	+27,12	0,0002*	0,6960	0,6280
	OrigMid-De-De	0,3900	+40,44	<0,0001*	0,7200	0,6340
	StemMid-De-De	0,4058	+46,13	<0,0001*	0,7040	0,6420
	OrigStemMid-De-De	**0,4135**	+48,90	<0,0001*	**0,7280**	**0,6520**
Englisch	Orig-En-En	0,2049			0,6000	0,4660
	Stem-En-En	0,2562	+25,04	0,0018*	0,6160	0,5540
	Mid-En-En	0,2330	+13,71	0,6528	0,5280	0,5020
	OrigStem-En-En	0,2491	+21,57	0,0002*	0,6320	0,5440
	OrigMid-En-En	0,2695	+31,53	<0,0001*	0,6080	0,5700
	StemMid-En-En	**0,2885**	+40,80	<0,0001*	**0,6400**	**0,5860**
	OrigStemMid-En-En	0,2839	+38,56	<0,0001*	0,6320	0,5660

3.5.4 GIRT - Ergebnisse

In Tabelle 3.6 sind die Ergebnisse der deutschen und englischen Testläufe für die GIRT-Kollektion dargestellt. Abbildungen 3.8 und 3.9 stellen die Ergebnisse für MAP und verschiedene Präzisionswerte in beiden untersuchten Sprachen grafisch dar.

Der beste Testlauf bezüglich MAP ist die Kombination aus allen drei Anfrageversionen (*OridStemMid-De-De*) in der deutschen Sprache mit einem Wert von 0,4135 (+48,90% gegenüber der Baseline *Orig-De-De*). Die englischen Testläufe schneiden konstant niedriger ab als die deutschen, der beste englische Testlauf ist *StemMid-En-En* mit einer MAP von 0,2885 (+40,80% gegenüber der Baseline *Orig-En-En*). Obwohl MorphoSaurus nicht auf die sozialwissenschaftliche Domäne trainiert ist, sind seine MAP-Werte mit 0,2846 gegenüber 0,2777 im Deutschen und 0,233 gegenüber 0,2049 im Englischen der Baseline ebenbürtig. Das Verfahren mit dem

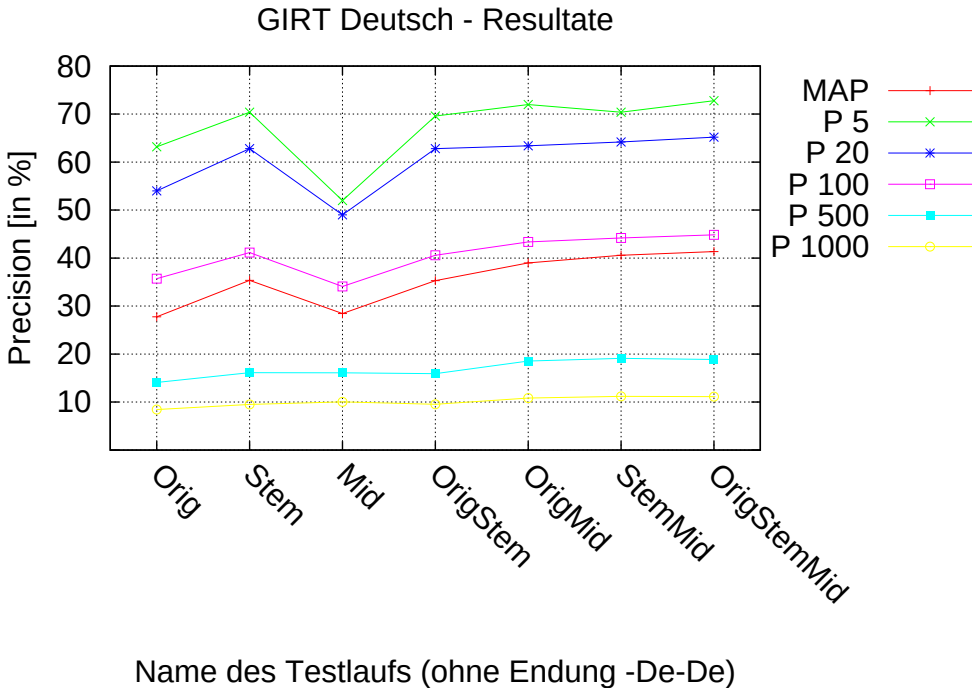

Abbildung 3.8: MAP sowie Precision-Ergebnisse der deutschen Testläufe bei verschiedenen Cut-Off-Punkten.

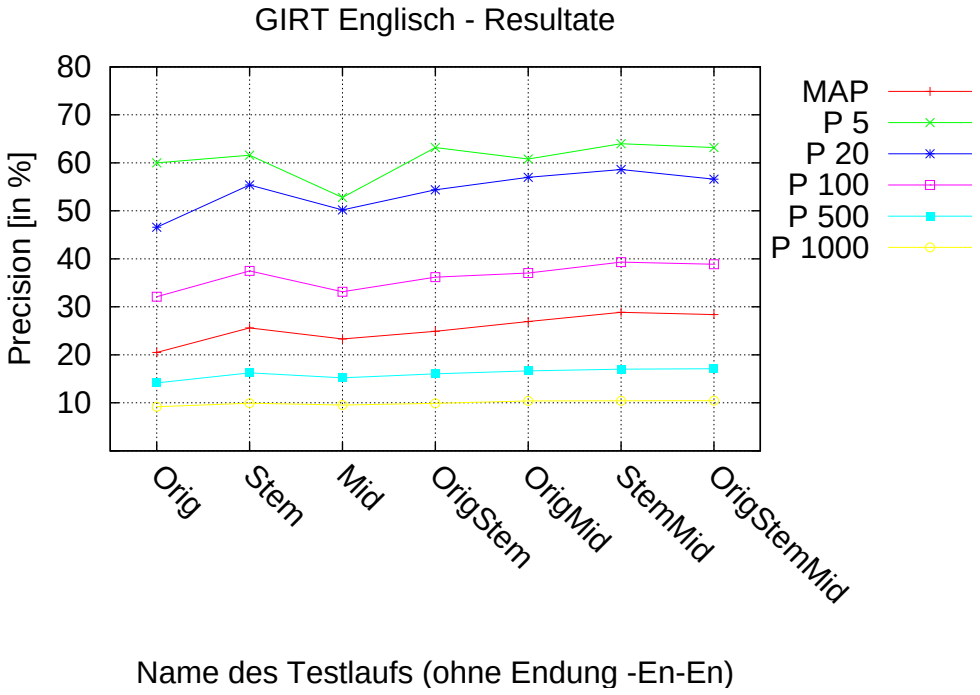

Abbildung 3.9: MAP sowie Precision-Ergebnisse der englischen GIRT-Testläufe bei verschiedenen Cut-Off-Punkten.

Porter-Stemmer ist allerdings mit 0,3535 im Deutschen und 0,2562 im Englischen beiden anderen Verfahren deutlich und statistisch signifikant überlegen.

Bei den P5 und P20 Werten schneidet das MorphoSaurus-System in der sozialwissenschaftlichen Domäne am schlechtesten ab. Der P5-Wert von *Mid-De-De* liegt mit 0,5200 deutlich unter *Stem-De-De* mit 0,7040. Auch im Englischen ist es mit 0,5280 gegenüber 0,6160 unterlegen. Die kombinierten Verfahren schneiden wie in den anderen Kollektionen am besten ab, die höchsten Werte erzielen im Deutschen *OrigStemMid* mit 0,7280 und im Englischen *StemMid* mit 0,6400.

3.6 Diskussion der Ergebnisse und Ausblick

3.6.1 Singuläre Testläufe

Die Diskussion beginnt mit den Ergebnissen der singulären Testläufe *Orig*, *Stem* und *Mid*. Zunächst einmal fällt auf, dass sich die drei Verfahren bis auf die Testläufe mit der GIRT-Kollektion nicht statistisch signifikant voneinander unterscheiden. Um die Ursache hierfür herauszufinden, wurde eine detaillierte Fehleranalyse durchgeführt, in der die Abweichungen von *Mid* und *Stem* gegenüber der Baseline *Orig* für alle Testkollektion auf Ebene der einzelnen Fragen ermittelt wurde. Die Abbildungen 3.10, 3.11, 3.12 und 3.13 geben an, für welche Testanfragen die Testläufe *Stem* und *Mid* besser oder schlechter abschneiden als die Baseline *Orig*. Nach unten weichen insbesondere die Fragen 9 und 86 aus der OHSUMED-Kollektion, die Fragen 19 und 28 aus der englischen ImageClef-Kollektion, die Fragen 3, 8 und 22 aus dem deutschen GIRT-Korpus sowie die Frage 9 aus dem englischen GIRT-Korpus ab. Positiv ragen die Fragen 1, 51 und 98 aus der OHSUMED-Kollektion, die Fragen 13 und 30 aus ImageCLEF, die Fragen 9 und 15 aus dem deutschen GIRT-Korpus sowie die Frage 4 aus dem englischen GIRT-Korpus heraus. In dieser Diskussion werden insbesondere die Abweichungen der *Mid*-Testläufe nach unten näher betrachtet:

Bei den meisten Abweichungen wird deutlich, dass die Performanz der MorphoSaurus-Suche deutlich sinkt, wenn in den MorphoSaurus-Lexika zu "unspezifische Äquivalenzklassen" definiert sind. "Unspezifische Äquivalenzklassen" bedeutet, dass eine ganze Reihe von Originalwörtern in einer gemeinsamen Äquivalenzklasse abgebildet werden. Beispiele hierfür werden nun beginnend mit dem OHSUMED-Korpus aufgezeigt:

Für Frage 9 *"Effectiveness of gallium therapy for hypercalcemia"* (*"Effektivität der Gallium-Therapie für Hyperkalzämie"*) werden im Gold-Standard insgesamt nur sechs Dokumente als relevant beurteilt. Die Original-Version findet fünf von sechs

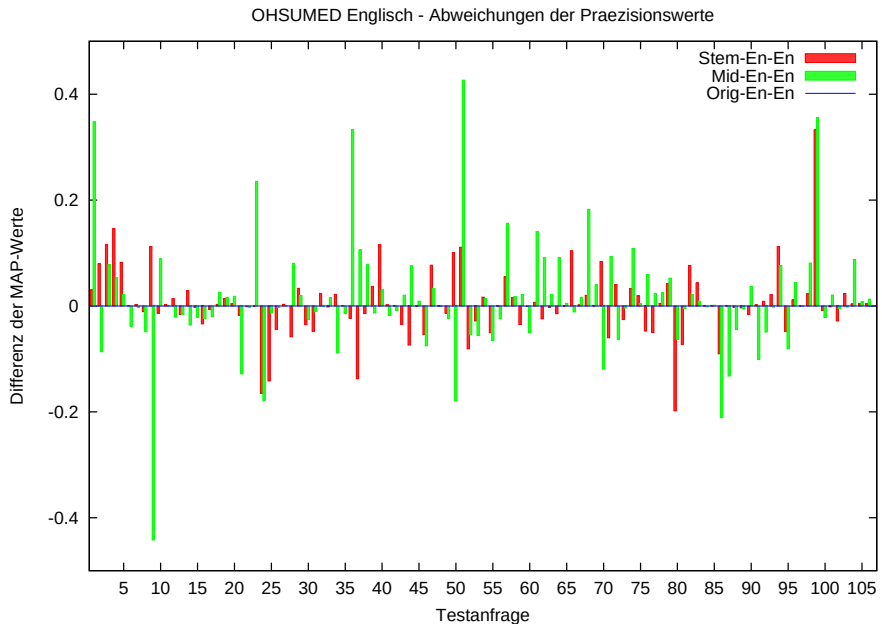

Abbildung 3.10: Abweichungen der MAP-Werte pro Testanfrage für OHSUMED.

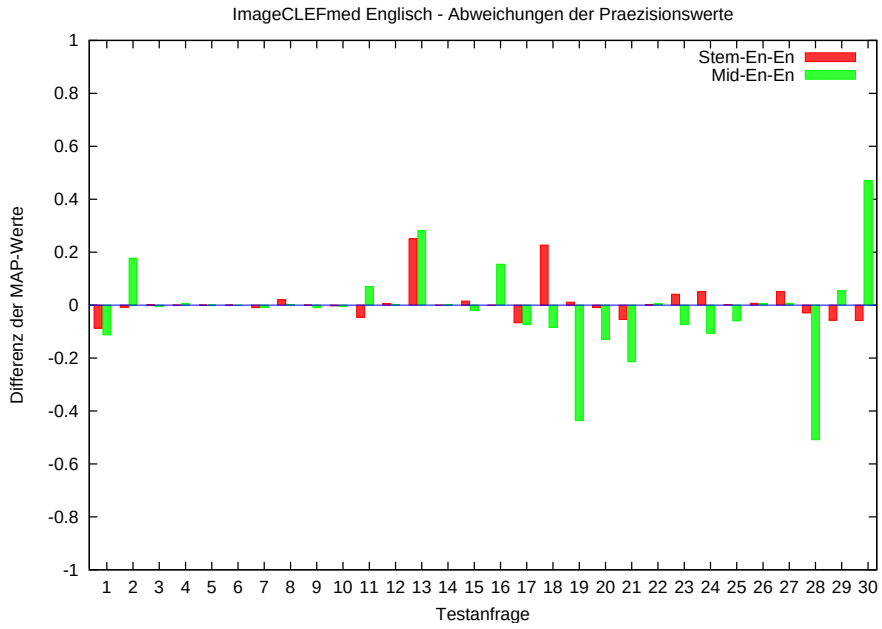

Abbildung 3.11: Abweichungen der MAP-Werte pro Testanfrage für ImageCLEF. Die Map-Werte von *Orig-En-En* gelten als Baseline, in Balken ausgedrückt sind die Differenzen der Testläufe *Stem-En-En* und *Mid-En-En*.

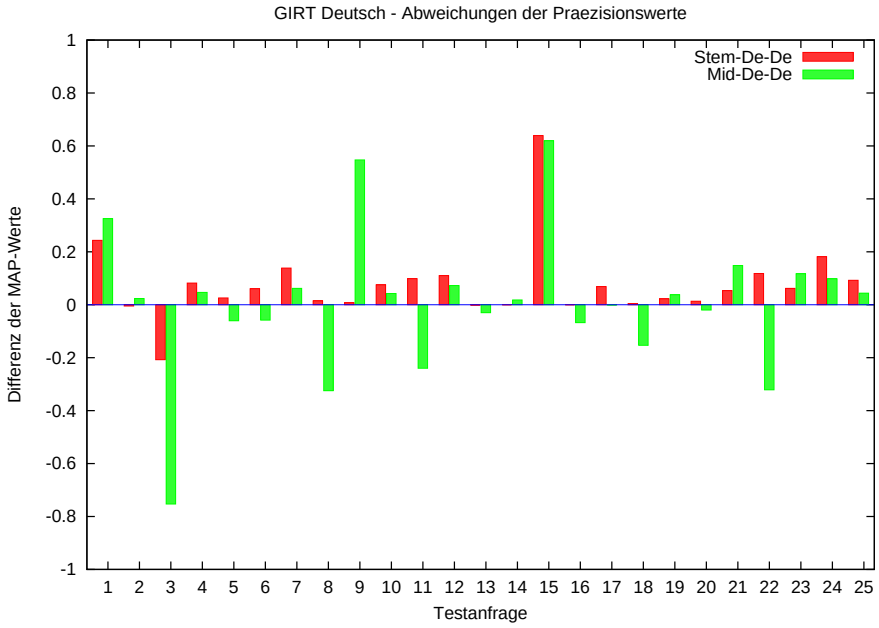

Abbildung 3.12: Abweichungen der MAP-Werte pro Testanfrage im deutschen GIRT-Korpus.

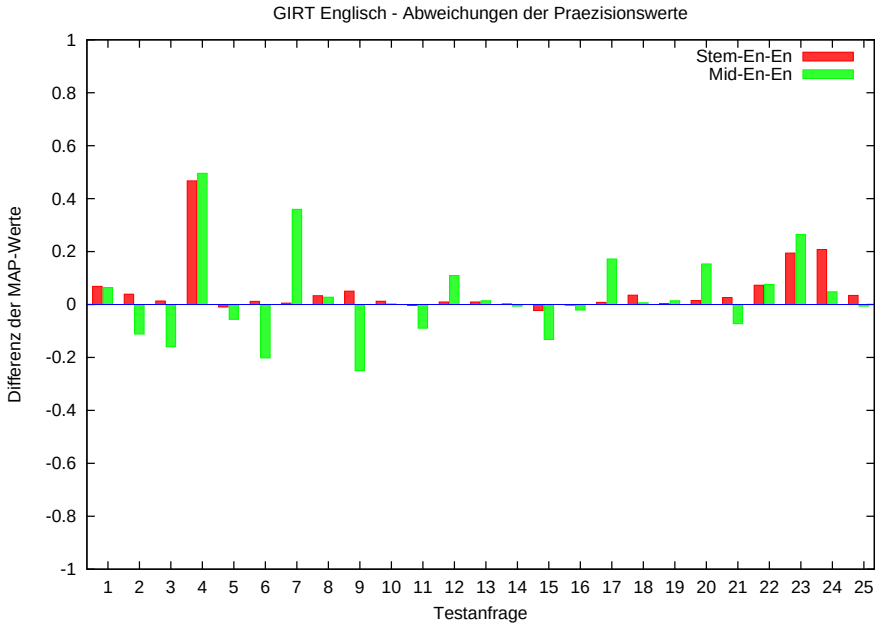

Abbildung 3.13: Abweichungen der MAP-Werte pro Testanfrage im englischen GIRT-Korpus.

relevanten Treffern, drei davon werden innerhalb der ersten fünf Dokumente aufgeführt. Die MID-Version findet ebenfalls fünf von sechs relevanten Treffern, allerdings werden diese nicht soweit vorne wie in der Original-Version einsortiert. Die MID-Version dieser Anfrage lautet *"#effect #gallium #treat #hyper #calc #blood"*. Durch die Auftrennung in viele kleinere Einheiten werden durch Lucene deutlich mehr Treffer gefunden, darunter befinden sich relevante und nicht relevante Treffer. Einige der nicht relevanten Treffer werden auch vor die relevanten Dokumente eingeordnet, so dass die relevanten Dokumente erst auf den hinteren Plätzen zu finden sind.

Die Frage 86 aus diesem Korpus lautet: *"Is increased thyroid stimulating hormone associated with sick euthyroid syndrome?" ("Ist erhöhtes Thyroid-stimulierendes Hormon verknüpft mit dem Euthyroid-Sick-Syndrom*[16]*")*. Der Terminus *"sick euthyroid syndrom"* ist dabei ein feststehender medizinischer Begriff für eine Schilddrüsenerkrankung. Da dieser Ausdruck jedoch nicht in dem MorphoSaurus-Lexikon enthalten ist, wird er in die relativ unspezifischen Äquivalenzklassen *"#path #ordinary #thyroid #syndrom"* überführt. Auf die Äquivalenzklasse *"#path"* werden Wörter wie *"disease, -ose, -pathy, sick, illness etc."* abgebildet, auf die Äquivalenzklasse *"#ordinary"* fallen Wörter wie *"norm, normal, ordinary, eu-"*. Damit werden auch viele nicht relevante Dokumente gefunden, beispielsweise das Dokument mit folgender Textpassage *"[...] Schilddrüsenerkrankung mit Werten, die nach einiger Zeit wieder in den Normalzustand übergingen.[...]"*. In diesem Ausschnitt erscheinen die Wort(-teile) *"Schilddrüse, Erkrankung"* und *"Normal"* und sind damit für die MID-Version relevant, dennoch sind diese Dokumente im Goldstandard korrekterweise nicht als relevant beurteilt worden, da sie nicht von dem *Sick-Euthyroid-Syndrom* handeln.

Als nächstes wird Frage 19 aus der ImageCLEF-Kollektion betrachtet: *"Show me images with multinucleated giant cells"*. Der Ausdruck *"Multinukleare Riesenzelle"* ist wiederum ein feststehender medizinischer Terminus. Dieser ist nicht im MorphoSaurus-Lexikon enthalten und wird daher auf *"#multi #nucleus #big #cell"* abgebildet. In der Äquivalenzklasse *"#big"* sind Subwörter wie *"big, huge, giant, tall, great, enorm"* enthalten. *"#nucleus"* enthält neben *"nucleus, nuclear"* auch das englische Wort *"core"*. Die unspezifische Definition dieser Äquivalenzklassen führt zu einigen falsch positiven Treffern.

Die Frage 28 (*"Show me microscopic images showing parvovirus infection"* (*"Zeige mir mikroskopische Bilder einer Parvovirus-Infektion"*)) aus der ImageCLEF-

[16]Sowohl *Euthyroid-Sick-Syndrom* als auch *Sick-Euthyroid-Syndrom* erscheint in der Literatur.

Kollektion wird vom MorphoSaurus-System in die Äquivalenzklassen *"#tiny #scope #tiny #virus #infection"* überführt, ein Eintrag für das Virus mit dem Namen *"Parvovirus"* fehlt im MorphoSaurus-Lexikon. Dadurch werden beispielsweise Dokumente als relevant gewertet, die Ausschnitte wie *"[...] Viruses cause tiny lesions [...]"* (*"Viren verursachen winzige Läsionen"*) beinhalten. Dies senkt die Performanz der MorphoSaurus-Suche.

In der deutschen GIRT-Kollektion werden die Frage 3 und 8 betrachtet. Frage 3 (*"Kinderlosigkeit in Deutschland"*) wird durch MorphoSaurus zerlegt in *"#child #no #german"*. Da auf die Äquivalenzklasse *"#no"* zahlreiche Subwörter wie *"-los"* (Bsp. *"kinderlos"*), *"kein"* (Bsp. *"kein Anzeichen für"*), *"des-"* (Bsp. *"Desinteresse"*) oder *"a-"* (Bsp. *"Anomalie"*) abgebildet werden, werden in dieser Anfrage viele falsch positive Treffer, d.h. Treffer, die die MorphoSaurus-Suche irrtümlich als relevant einstuft, gefunden.

Frage 8 (*"Rückwanderung und Transmigration"*) erhält die Äquivalenzklassen *"#rear #wander #across #migr"* zugewiesen. Auf *"#rear"* werden Subwörter wie *"rück, hinten, hinter, nach"* abgebildet, auf *"#across"* Subwörter wie *"trans-, hinüber, quer"*. Damit sind zwei der vier Äquivalenzklassen zu ungenau, um exakte Treffer zu erzielen.

Bei Frage 22 (*"Ausländer in der Grundschule"*) ist das Ergebnis deshalb deutlich schlechter, da bisher weder *"Ausländer"* noch *"Grundschule"* im MorphoSaurus-Lexikon definiert wurden und damit in die Äquivalenzklassen *"#ex #country"* und *"#base #school"* zerlegt werden. Sowohl *"#ex"* als auch *"#base"* stellen dabei sehr unspezifische Klassen dar, die zu vielen falsch positiven Treffern führen.

Die Frage 4 aus dem englischen GIRT-Korpus heißt *"role of the father"* (*"Rolle des Vaters"*). Im MorphoSaurus-Lexikon ist *"father"* zusammen mit *"patriarch"* in einer Äquivalenzklasse. In der GIRT-Kollektion existieren beispielsweise auch Dokumente über *"role of the maternal patriarch"*, die im Gold-Standard korrekterweise als nicht relevant eingeschätzt wurden, aber von der MorphoSaurus-Suche als relevant beurteilt wurden.

3.6.2 Kombinierte Testläufe

Ist der Nutzen von MorphoSaurus an den singulären Testläufen nicht überzeugend abzulesen, so ist aus Sicht des MorphoSaurus-Systems dennoch die Tatsache erfreulich, dass alle kombinierten Verfahren mit MorphoSaurus besser abschneiden als die Verfahren ohne dieses System. In allen Testsets sind die Kombinationen mit MorphoSaurus bezüglich des MAP-Wertes 15-50% besser als die Baseline. Der

größte Vorteil des MorphoSaurus-Systems zeigt sich überraschenderweise in der sozialwissenschaftlichen Domäne des GIRT-Korpus für die deutsche Sprache, in der der Testlauf *OrigStemMid-De-De* 48,9% besser als die Baseline *Orig* und 17% besser als der Testlauf mit dem Porter-Stemmer *Stem* abschneidet. Dabei sind die Vorteile, die sich durch die morpho-semantische Analyse mit dem MorphoSaurus-System ergeben, in der deutschen Sprache wie erwartet am ersichtlichsten. Für die englische Sprache ist das Ausmaß der Verbesserung bezüglich MAP mit 15% (ImageCLEF) bis 39% (GIRT) gegenüber der Baseline etwas moderater.

3.6.3 Fazit

Die Ergebnisse verdeutlichen, dass der Ansatz des MorphoSaurus-Systems zur morpho-semantischen Analyse allein derzeit nicht zu einer Verbesserung der Retrieval-Ergebnisse führt. Der Erfolg des Systems hängt offensichtlich ganz entscheidend von der Qualität und der Abdeckung der lexikalischen Ressourcen des MorphoSaurus-Systems ab. Fehlen in den Ressourcen Einträge oder sind die bestehenden Einträge zu unspezifisch definiert, führt das zu Einbußen in den Retrieval-Ergebnissen. Ursprünglich wurde das MorphoSaurus-System als ein System betrachtet, mit dem aufgrund der Berücksichtigung zahlreicher linguistischer Variationen die "Nadel im Heuhaufen" gefunden werden kann. Diese Aussage erweist sich nach den hier vorgestellten Ergebnissen als nicht generell gültig, da durch unspezifische Äquivalenzklassen die Zahl der falsch positiven Treffer deutlich ansteigt, was den Nutzen, dass auch einige relevante Dokumente gefunden werden, deutlich überwiegt. Anders ausgedrückt, findet das MorphoSaurus-System zwar einige relevante Dokumente mehr als die Originalsuche oder die Stammform-basierte Suche, dafür aber auch viele nicht relevante Dokumente. Die guten Ergebnisse insbesondere für die OHSUMED-Kollektion sowie in einzelnen Anfragen in den anderen Kollektionen machen jedoch deutlich, dass in dem Verfahren zur morpho-semantischen Normalisierung durchaus einiges Potential steckt, welches es in der Zukunft herauszuarbeiten gilt. Dies bezieht sich in erster Linie auf die Arbeit an den lexikalischen Ressourcen des MorphoSaurus-Systems. Hier erscheint insbesondere ein Abweichen von der Strategie sinnvoll, alle Subwörter, die in irgendeinem Kontext synonym verwendet werden, in einer gemeinsamen Äquivalenzklasse zu definieren. Als Beispiel sei die bereits erwähnte Äquivalenzklasse *"#father"* genannt, in denen die Subwörter *"father"* und *"patriarch"* enthalten sind, welche nur selten synonym verwendet werden. Insbesondere einige sehr unspezifische Äquivalenzklassen wie *"#multi, #tiny, #across, #rear"* sollten darüber hinaus überarbeitet werden und in mehrere, enger

gefasste Äquivalenzklassen aufgeteilt werden.

Eine zweite Möglichkeit zur Verbesserung der MorphoSaurus-Ergebnisse könnte dadurch erreicht werden, dass die Nähebeziehung, die von Lucene für die Formulierung der Anfrage benutzt wird, ausgenutzt wird. Durch die Aufspaltung von (wenigen) Originalwörtern in (viele) Subwörter geht die Nähebeziehung zusammengesetzter Wörter aus der Originalanfrage zunächst verloren. Dies kann bei der Relevanzbeurteilung von Lucene negative Auswirkungen haben, wie sich beispielsweise in Frage 9 der OHSUMED-Kollektion zeigte. Die Nähebeziehung könnte in Lucene jedoch durch die *Distanzsuche* reproduziert werden. In bisherigen IR-Szenarien, in denen die Näheinformation berücksichtigt wurde, ergaben sich bisher widersprüchliche Ergebnisse über deren Nutzen [Markó et al.2005a, Markó2007]. Ausführliche, klärende Untersuchungen stehen noch aus.

Als sehr erfolgsversprechend erweisen sich die vorgestellten Kombinationsansätze. Sie verdeutlichen den Nutzen des MorphoSaurus für die monolinguale Textrecherche, da die Testläufe mit MorphoSaurus teilweise erheblich besser abschneiden als die Testläufe ohne dieses System.

In der Einleitung wurde MorphoSaurus als ein System vorgestellt, dass insbesondere in stark flektierenden und agglutinierenden Sprachen wie dem Deutschen, Holländischen oder Schwedischen zu erheblichen Verbesserungen der Performanz von NLP-Systemen führen kann. In dem nicht-medizinischen deutschen GIRT-Korpus hat sich dies für den kombinierten Testlauf *OrigStemMid* bereits eindrucksvoll bestätigt. Für die medizinische Domäne, auf die das MorphoSaurus-System trainiert ist, stehen leider nicht in ausreichend großer Zahl Testdokumente und Gold-Standards zur Verfügung. Das deutsche ImageCLEF-Subkorpus erwies sich hierfür als zu klein. Daher konnte das MorphoSaurus-Systems lediglich für die englische Sprache ausreichend überprüft werden, in der der Nutzen erwartungsgemäß moderater ausfällt, da im Englischen nur wenige linguistische Variationen existieren. Eine groß angelegte Evaluation auf der deutschen Medizindomäne steht noch aus.

Die Frage 87 aus dem OSHUMED-Korpus und Frage 19 aus dem englischen ImageCLEF-Korpus, in denen die Ausdrücke *"Euthyroid Sick Syndrome"* und *"Multinucleated Giant Cell"* vom MorphoSaurus-System nicht erkannt wurden, verdeutlichen, dass insbesondere für Mehrwortausdrücke zusätzliche Strategien im MorphoSaurus-System entwickelt werden müssen. Eine Aufnahme aller Mehrwortausdrücke in die MorphoSaurus-Lexika erscheint sowohl aus zeitlichen wie auch aus laufzeitlichen Gründen als nicht sinnvoll. Hier sollten insbesondere bestehende Quellen wie der UMLS, in dem bereits zahlreiche Mehrwortausdrücke vor-

handen sind, in irgendeiner Form in das MorphoSaurus-System integriert werden (siehe hierzu auch das Kapitel 2.7). Sinnvolle Lösungen werden derzeit diskutiert.

3.7 Verwandte Arbeiten

Dem Autor ist kein anderes IR-System bekannt, welches ähnlich wie das MorphoSaurus-System Wörter in ihre Wortbestandteile zerlegt und basierend auf den Wortbestandteilen Synonymiebeziehungen definiert und Sinnambiguitäten auflöst. Will man das MorphoSaurus-System bezüglich seiner "linguistischen Komplexität" einordnen und hinsichtlich seiner Funktionalitäten mit anderen Ansätzen vergleichen, so nimmt es wohl eine Stellung zwischen einfachen Stemming-Verfahren und aufwändigen NLP basierten Retrievalarchitekturen ein. Die wichtigsten Arbeiten, die zum MorphoSaurus-System gewisse Ähnlichkeiten aufweisen und im Information Retrieval eingesetzt werden, werden nun näher betrachtet.

Bei den Verfahren zur Stammformbildung werden regelbasierte Verfahren [Lovins1968, Porter1980, Savoy2006], lexikonbasierte Verfahren [Tomlinson2004] sowie statistische Verfahren [Nunzio et al.2004] unterschieden. Eine Sonderstellung nehmen *N-Gram* basierte Verfahren ein [McNamee & Mayfield2004], bei denen Wörter nicht in ihre Stammformen, sondern in N-Gramme (typischerweise 3-Gramme bis 5-Gramme) zerlegt werden. Für zahlreiche Sprachen wurde nachgewiesen, dass sich die Normalisierung von Wörtern positiv auf die IR-Ergebnisse auswirken. Studien existieren unter anderem für Slowenisch [Popovič & Willett1992], Italienisch [Sheridan & Ballerini1996], Holländisch [Kraaij & Pohlmann1996], Deutsch [Moulinier et al.2001, Braschler & Ripplinger2004], Finnisch [Airio2006]. [Tomlinson2001] hat Stemming für neun Sprachen untersucht. Er fand heraus, dass für das Deutsche und das Finnische Lexikon-basierte Stemming-Verfahren dem regelbasierten Porter-Stemmer überlegen sind. Für die anderen Sprachen gab es keine signifikanten Unterschiede. In Zahlen ausgedrückt lagen die Performanzgewinne gegenüber der Baseline zwischen 1,3% (Englisch) und 25,2% (Finnisch). Da das Lexikon-basierte Verfahren ein kommerzielles Werkzeug ist, liegen keine Angaben über den Umfang der lexikalischen Ressourcen vor. [Nunzio et al.2004] untersuchte ein statistisches Stemming-Verfahren für fünf Sprachen und zeigte, dass die Vorteile dieses Verfahrens bezüglich der Retrieval-Performanz zwischen den verschiedenen Sprachen variieren. Insgesamt waren die statistischen Verfahren dem regelbasierten Porter-Stemmer jedoch leicht unterlegen.

[Hollink et al.2004] untersuchten für acht Sprachen Stammformbildung, Dekomposition und N-Gram-Techniken und fanden, dass für alle Sprachen außer dem Engli-

schen linguistisch motivierte Ansätze besser abschneiden als die Baseline. Allerdings gab es kein einheitlich bestes Verfahren für alle Sprachen. Spanisch war die einzige Sprache, welche bei der Stammformbildung signifikant besser abschnitt als bei N-Gramm-Techniken. [Savoy2006] schließlich entwickelte für vier europäische Sprachen einfache regelbasierte Stemming-Techniken. In allen Sprachen wurden signifikante Verbesserungen der Performanz gegenüber der Baseline erzielt.

Techniken zur Dekomposition sind in den Ansätzen von [Moulinier et al.2001, Kraaij & Pohlmann1996, Tomlinson2004, Hollink et al.2004] und [Braschler & Ripplinger2004] durchgeführt worden. Die Performanzgewinne in den Tests liegen zwischen 17% und 54%. [Braschler & Ripplinger2004] konnte zeigen, dass sich der Effekt durch Dekomposition zusammengesetzter Wörter für das Deutsche positiver bemerkbar macht als Stammformbildung (34% vs. 16% für kurze Anfragen und 28% vs. 9% für lange Anfragen).

Keines dieser in den Studien angesprochenen Verfahren beherrscht den Umgang mit syntaktischen Variationen, die im MorphoSaurus-System in Form von Mehrwortausdrücken teilweise unterstützt werden. Syntaktische Variationen werden typischerweise von Phrasen-basierten IR-Systemen behandelt. Arbeiten hierzu existieren beispielsweise von [Fagan1989, Jacquemin et al.1997, Mitra et al.1997, Wessel Kraaij1998, Koster1999, Smeaton1999, Arampatzis et al.2000]. Der Erfolg dieser Verfahren im Information Retrieval ist allerdings moderat [Brants2003]. Da diese Verfahren nur schwer mit dem MorphoSaurus vergleichbar sind und eine Erläuterung dieser Verfahren über den Rahmen dieser Arbeit hinausginge, werden sie an dieser Stelle nicht weiter beschrieben.

Einige Arbeiten existieren zur semantischen Erweiterung von Anfragen in Form von Synonymen auf Wortebene. Häufig wird bei diesen Verfahren das allgemeinsprachliche Vokabular WordNet [Fellbaum1998] oder im medizinischen Bereich der UMLS [Umls2005b] eingesetzt. [Voorhees1994] zeigte beispielsweise für kurze Anfragen, dass durch eine Erweiterung dieser Anfragen mit auf WordNet basierten Termen die Retrievalergebnisse signifikant gesteigert werden konnten. Für lange Anfragen konnte dies jedoch nicht nachgewiesen werden. In Arbeiten von [Gonzalo et al.1998] führte eine auf WordNet basierte Anfrageerweiterung nur dann zu Performanzgewinnen, wenn die Originalanfragen zuvor disambiguiert wurden. [Smeaton et al.1995] erweiterten spezifische Terme mit übergeordneten und untergeordneten Begriffen sowie mit Synonymen aus WordNet. Die Ergebnisse waren jedoch enttäuschend.

[Aronson & Rindflesch1997] führten Experimente basierend auf Testdaten von [Hersh et al.1994b] durch und verglichen die Ergebnisse mit Studien von

[Srinivasan1996b], die ähnliche Experimente durchführte. Als Werkzeug zur Anfrageerweiterung wurde MetaMap eingesetzt [Aronson et al.1994], welches Freitext automatisch auf Terme des UMLS abbildet. Im Vergleich zu den nicht-erweiterten Anfragen erzielten die Verfahren minimale Performanzgewinne von 4,4% (Aronson) bzw. 2,2% (Srinivasan).

[Hersh et al.2000] führten Tests auf der OHSUMED-Kollektion durch und ordneten jeder Frage manuell entsprechende UMLS-Terme zu. Diese erweiterten Anfragen wurden dem SAPHIRE-System [Hersh & Leone1995] zur weiteren Ergänzung von UMLS-Termen übergeben. Insgesamt standen somit 106 Testanfragen mit 298 Termen zur Verfügung. In verschiedenen Testläufen wurden nun Anfragen mit und ohne Termerweiterung sowie zusätzlich mit hierarchischen Erweiterungen aus dem UMLS-Thesaurus getestet. In 38,6% der erweiterten Anfragen (ohne hierarchische Erweiterung) sowie in 29,7% der Anfragen mit hierarchischer Erweiterung konnte eine Steigerung der Performanz verzeichnet werden, insgesamt betrachtet kam es jedoch zu einem Performanzverlust.

Kapitel 4

Textkategorisierung mit MorphoSaurus

4.1 Einleitung

Die Kategorisierung von Dokumenten, bei der Texte in vordefinierte Gruppen eingeordnet werden und somit zusätzliche Informationen (Metadaten) hinzugefügt werden, ist eine wichtige Aufgabe bei der Dokumentenrecherche. Durch Angabe zusätzlicher Informationen werden Dokumente näher charakterisiert, so dass Suchende diese Dokumente leichter und genauer finden können. Durch die Eingruppierung in inhaltlich verwandte Bereiche werden die Suchmöglichkeiten um die Verwendung von Synonymen, Ober- und Unterbegrifflichkeiten und weiteren Merkmalen erweitert. Das Abbilden auf kontrollierte Terminologien ist darüber hinaus entscheidend für die semantische Operabilität zwischen Krankenhäusern sowie für den länderübergreifenden Austausch medizinischer Daten. So schrieb das Direktorium der amerikanischen Vereinigung für medizinische Informatik bereits im Jahre 1994: *"...standards for codes/terminology are an essential requirement for a computer-stored medical record that spans more than one provider's domain."* [AMIA1994]. In mehreren Ländern wird dementsprechend mit SNOMED CT [Snomed CT2006] derzeit eine Referenzterminologie eingeführt.

Metadaten können in Schlagwörter und Attribute unterteilt werden. Schlagwörter repräsentieren den Inhalt der Dokumente. Attribute enthalten Informationen über den Typ des Dokumentes (z.B. Zeitschriftenartikel), sein Erscheinungsjahr oder auch den Ort der Veröffentlichung. Der Prozess des Hinzufügens von Metadaten wird als *Kategorisierung, Indexierung, Tagging* oder *Verschlagwortung* bezeichnet. Viele bibliographische Datenbanken bedienen sich bei der Auswahl der

Schlagwörter eines *kontrollierten Vokabulares* oder eines *Thesaurus*, welche für derartige Zwecke erstellt werden. Die Verschlagwortung der Dokumente mit Hilfe der Einträge aus einem solchen Vokabular kann manuell oder automatisch erfolgen.

Die manuelle Indexierung medizinischer Dokumente reicht weit vor das Computerzeitalter zurück. 1879 erstellte John Shaw Billings den ersten Index für medizinische Literatur, *Index Medicus*, der Dokumente nach ihrem Inhalt klassifizierte. Im letzten Jahrzehnt wurde der Index Medicus durch die elektronische Version MEDLINE abgelöst. MEDLINE wird von der U.S. National Library of Medicine (NLM) gepflegt und ist unter anderem durch das Suchinterface PubMed[1] zugänglich. Die Verschlagwortung der Dokumente in MEDLINE mit dem MeSH-Vokabular ist nach wie vor eine intellektuelle Expertenarbeit, die zeit- und kostenintensiv ist. Die NLM beschäftigte bereits in den Neunzigern 44 Vollzeitindexierer und gab über 2 Millionen Dollar für die Indexierung ihrer Dokumente aus. Seit dieser Zeit hat sich die Zahl der Dokumente, die jährlich zu MEDLINE hinzugefügt wird, mit derzeit 350.000 Dokumenten vervierfacht.

Bei der manuellen Indexierung entstehen neben den Kosten eine Reihe weiterer Probleme. Dazu gehören Inkonsistenzen, die dadurch entstehen, dass verschiedene Indexierer gleiche Dokumente unterschiedlich verschlagworten. [Funk & Reid1983] evaluierten die Konsistenz bei der MEDLINE-Indexierung mit Hilfe von 760 Dokumenten, die versehentlich zwei Mal indexiert wurden. Die Übereinstimmung bei den eigentlichen Schlagwörtern (MeSH-Headings) der MEDLINE-Indexierung betrug lediglich 48,2% (für Schlagwörter) bzw. 61,6% (für zentrale Schlagwörter). Tabelle 4.1 gibt einen genauen Überblick über die Konsistenzraten.

[Crain1987] identifizieren drei Hauptgründe, die zu Inkonsistenzen führen:

1. Vorerfahrungen bei der MEDLINE-Indexierung - manche Indexierer besitzen "Lieblings-"Schlagwörter, die ohne große Reflektion häufig verwendet werden.

2. Eigene Regeln bei der Einschätzung der Wichtigkeit von Schlagwörtern - Indexierer tendieren dazu, Schlagwörter als wichtig einzustufen, die ihnen unbekannt sind.

3. Unterschiedliche Interpretation der Indexierungsregeln, die den Indexierern von der NLM vorgegeben werden.

In Anbetracht der ansteigenden Informationsflut sowie der schwindenden finanziellen Ressourcen stellt die Entwicklung automatisierter Verfahren zur Verschlagwortung medizinischer Dokumente ein dringliches Ziel bei der Dokumentenrecherche

[1] Zugriff über http://www.pubmed.org, eingesehen im Februar 2007

Tabelle 4.1: Konsistenz der MEDLINE-Indexierung nach MeSH-Kategorie (aus [Funk & Reid1983])

Kategorie	Konsistenz (in %)
Check tags	74,7
Zentrale Headings	61,6
Geographie	56,6
Zentrale Subheadings	54,9
Subheadings	48,7
Headings	48,2
Zentrale Heading/Subheading Kombination	43,1
Heading/Subheading Kombination	33,8

dar. Die Herausforderungen an solche Systeme fasst [Hersh2002] unter folgenden Punkten zusammen:

- **Synonymie**: Dokumente über "Hypertonie" sollten gleich behandelt werden wie Dokumente über "Bluthochdruck".
- **Polysemie**: Manche Wörter wie das Wort "Bruch" ("Fraktur" bzw. "Hernie") sind mehrdeutig und sollten disambiguiert werden.
- **Kontext**: Manche Wörter haben in gemeinsamen Kontext eine andere Bedeutung. *"High"*, *"blood"* und *"pressure"* zum Beispiel nehmen eine zusätzliche Bedeutung ein, wenn sie gemeinsam auftreten.
- **Morphologie**: Wörter können mit Suffixen versehen sein, welche die inhaltliche Bedeutung des Wortes nicht ändern und welche daher nicht die Indexierung beeinflussen sollten.
- **Granularität**: Anfragen und Dokumente besitzen häufig unterschiedliche Granularitätslevel. Beispielsweise wird eine Anfrage über *"Antibiotika"* gestellt, relevante Dokumente beschreiben jedoch bestimmte Antibiotika wie "Penizilline".

Die Indexierer der NLM verwenden, soweit das Dokument verfügbar ist, in der Regel den gesamten Artikel für die Verschlagwortung des MeSH-Vokabulars. Bei ausländischen Artikeln sind die Indexierer darauf angewiesen, dass zumindest eine englischsprachige Zusammenfassung des Artikels mitgeliefert wird. In jedem Fall ist

die Indexierung dieser Dokumente erheblich erschwert. Auf Grundlage von MorphoSaurus ist ein System entwickelt worden, welches Dokumente verschiedener Sprachen mit Schlagwörtern eines kontrollierten Vokabulars versieht und dabei viele der genannten Herausforderungen berücksichtigt. Für die in MorphoSaurus unterstützten Sprachen ist eine sprachübergreifende Indexierung möglich, d.h. es können beispielsweise deutsche Dokumente mit englischen Schlagwörtern annotiert werden. Die Evaluierung unseres Systems erfolgt mit Hilfe von Abstracts in verschiedenen Sprachen, denen Schlagwörter aus dem MeSH-Vokabular zugewiesen werden.

4.2 Das MeSH Vokabular

Der MeSH (Medical Subject Headings) [MeSH2005] wurde an der National Library of Medicine (NLM)[2] entwickelt und ursprünglich zur Indexierung des *Index Medicus* und später der MEDLINE-Datenbank eingesetzt. Mittlerweile ist der MeSH in verschiedene Sprachen übersetzt und wird von zahlreichen anderen Organisationen verwendet. Die englische Version wird von der NLM jährlich aktualisiert, und liegt als gedruckte Ausgabe sowie zum kostenfreien Download als Datenträgerversion vor. Jeder Deskriptor im MeSH enthält u.a. folgende Einträge:

- Main Headings (Hauptschlagwörter, Vorzugsbezeichnungen, Controlled Terms, CTs). In der Version von 2006 enthält der englische MeSH 23.885 sogenannte *Main Headings*.
- Entry terms (Synonyme, alternative Bezeichnungen, ETs), ETs haben Querverweise auf ein Main Heading.
- Erlaubte Subheadings (Qualifier) sind Zusatzbezeichnungen, die ein Main Heading näher spezifizieren und sich mit einem Schrägstrich an die vorangegangene Bezeichnung anschließen, z.B. /classification. Beispiel: “Taxonomie der Pilze” = fungi/cl.
- Eine Vielzahl von Querverweisen auf andere Deskriptoren, Anmerkungen, Definitionen der Deskriptoren, Hinweise auf die Verwendung sowie auf die “History” eines Deskriptors.

Der MeSH ist polyhierarchisch strukturiert. Er umfasst 15 Kategorien, die in Tabelle 4.2 abgebildet sind. Wie in der Tabelle dargestellt werden die Hauptkategorien jeweils mit einem Buchstaben bezeichnet. Die Kategorien gliedern sich über

[2] `www.nlm.nih.gov`, eingesehen im Februar 2007

Tabelle 4.2: Die 15 Hauptkategorien in MeSH

A. Anatomie	I. Anthroplogie, Erziehung, Soziologie
B. Organismen	und soziale Phänomene
C. Krankheiten	J. Technologie und Essen
D. Chemie und Arzneimittel	K. Sozialwissenschaften
E. Analyse, Diagnosen, Therapien	L. Informationswissenschaften
und Ausstattung	M. Personen
F. Psychiatrie und Psychologie	N. Gesundheitswesen
G. Biologie	Z. Geographische Lokalisationen
H. Physik	

mehrere Ebenen in Subkategorien auf. Eine Subkategorie kann dabei auch mehreren Oberkategorien zugeordnet werden, so dass die MeSH-Struktur weniger einen Baum als vielmehr ein Netz darstellt. Navigiert man in der Kategorie *A Anatomie* einige Ebenen nach unten, ergibt sich folgendes Bild auf die Hierarchie des MeSHs:

A Anatomy
A01 ... Body Regions
A01.047 ... Abdomen
A01.047.025 ... Abdominal Cavity
A01.047.025.600 ... Peritoneum
A01.047.025.600.225 ... Douglas Pouch

Peritoneum ist darüber hinaus in folgendem Ast enthalten:

A Anatomy
A10 ... Tissues
A10.615 ... Membranes
A10.615.789 ... Serous Membrane
A10.615.789.596 ... Peritoneum

Für die 23.885 Headings sind 148.063 Synonyme definiert. Einige davon sind *"print entry terms"*, welche in der Druckversion als Synonyme für das Heading angegeben werden. Die meisten jedoch sind *"nonprint entry terms"* und sind Variationen des Headings bezüglich Wortreihenfolge, Genus oder Schreibweise.

Tabelle 4.3: Die *Check Tags* von MeSH.

Animal	In Vitro
Case Report	Male
Comparative Study	Pregnancy
English Abstract	Support, Non-U.S. Gov't
Female	Support, U.S. Gov't, Non-P.H.S.
Human	Support, U.S. Gov't, P.H.S.

Tabelle 4.4: Die *Altersgruppen* in MeSH.

Infant, Newborn (1M)	Adult (19-44J)
Infant (2-24M)	Middle Age (45-64J)
Child, Preschool (2-5J)	Aged (65-79J)
Child (6-12J)	Support, Aged, 80 and over (80+J)
Adolescence (13-18J)	

Für eine exakte Indexierung der Dokumente sind die Headings bisweilen nicht ausreichend fein granular. Aus diesem Grund wurden weitere MeSH-Elemente eingeführt. Eines davon sind *Subheadings* oder *Qualifiers*, die zu den MeSH-Deskriptoren hinzugefügt werden und so die Bedeutung dieser Deskriptoren näher eingrenzen können. Insgesamt sind 82 Subheadings definiert, diese sind wie die MeSH-Deskriptoren hierarchisch geordnet. Für viele MeSH-Deskriptoren sind nur bestimmte Subheadings erlaubt, diese sind als *"Allowable Qualifiers"* für jeden MeSH-Deskriptor definiert.

Ein weiteres MesH-Element zur Indexierung von Artikeln sind die *Check Tags*. Diese Deskriptoren müssen zusätzlich zu den Headings jedem Artikel hinzugefügt werden, falls sie auf den Artikel zutreffen. Insgesamt stehen 12 Check Tags zur Verfügung, die in Tabelle 4.3 aufgelistet sind.

Altersgruppen sind eine weitere MeSH-Kategorie, die angewendet wird, wenn in den zu indexierenden Artikeln von Personen die Rede ist, wie es beispielsweise bei klinischen Tests auftritt. Altersgruppen werden nur in die Indexierung mit aufgenommen, wenn das Alter von Personen explizit in den Artikeln angegeben ist. Die vom MeSH verwendeten Altersgruppen sind in Tabelle 4.4 angegeben.

Geografische Angaben sind unter dem Buchstaben *Z* in der MeSH-Struktur enthalten. Diese Deskriptoren enthalten Kontinente, Länder, Regionen und weitere geografische Angaben.

Publication Types wurden 1991 als Erweiterung der früheren *Citation Types* eingeführt. Sie beziehen sich nicht auf den Inhalt eines Dokumentes, sondern charakterisieren die Art der Publikation näher. Dies kann beispielsweise ein Brief, ein historischer Artikel oder eine klinische Konferenz sein.

Neben dem eigentlichen MeSH stehen darüber hinaus die sogenannten *Supplementary Chemical Records* in einer Erweiterung des MeSHs zur Verfügung. Diese enthalten 164.502 Einträge über chemische Bezeichnungen.

Für das MeSH-Mapping wurde die englische Version 2006 des MeSH verwendet. In den Experimenten wurde sich auf die Indexierung von Dokumenten auf die Hauptkategorien des MeSH konzentriert. Es wurde nicht versucht, den MeSH Mainheadings zugehörige Subheadings zuzuordnen. Weiterhin wurde auf die Liste der Supplementary Chemical Records verzichtet, auch die Liste der Publication Types wurde nicht berücksichtigt. Mitberücksichtigt wurden hingegen die speziellen Deskriptoren *Check Tags, Age Groups* und *Geographics*. Diese Auswahl wird ebenso von der NLM für einige ihrer Experimente getroffen, die Verfahren zur automatischen Zuweisung von englischen Texten auf MeSH Deskriptoren entwickelt hat. Somit wird ein Vergleich unserer Ergebnisse mit denen der NLM ermöglicht.

Um eine effektive Zuordnung von MeSH-Termen zu Dokumenten mit dem MorphoSaurus-System zu ermöglichen, wurden zunächst einige Ergänzungen an den MorphoSaurus-Lexika durchgeführt. Dabei wurde mit Hilfe der MorphoSaurus-Analysetools überprüft, ob diejenigen Subwörter, aus denen ein MeSH-Term aufgebaut ist, bereits in den MorphoSaurus-Lexika enthalten sind. Falls dies nicht der Fall war, wurden die entsprechenden Subwörter in deutscher und englischer Sprache in die MorphoSaurus-Lexika aufgenommen. Insgesamt wurden somit 17.130 Einträge für die deutsche und englische Sprache ergänzt.

4.3 Die Indexierungsverfahren

Für die Indexierung von Dokumenten mit MeSH-Deskriptoren mit Hilfe des MorphoSaurus-Systems wurden verschiedene Verfahren entwickelt, die englische, deutsche, portugiesische, spanische und französische Dokumente mit englischen MeSH-Deskriptoren verschlagworten. Das erste Verfahren ist ein regelbasierter Ansatz, der ohne Trainingsdaten auskommt. Anhand einer Menge von Bewertungsfaktoren, die sich auf den Kontext innerhalb des Dokumentes stützen, werden relevante MeSH-

Terme ermittelt. Das zweite Verfahren ist ein statistischer Ansatz, der Trainingsdaten in Form bereits mit MeSH annotierter Dokumente benötigt. Diese wurden in verschiedenen Sprachen aus dem Web extrahiert. Mit Hilfe dieser Trainingsdaten werden bei Eingabe eines zu indexierenden Dokumentes diejenigen MeSH-Terme ermittelt, die bereits früher in einem ähnlichen Kontext verwendet wurden und daher mit einer gewissen Wahrscheinlichkeit relevante Deskriptoren darstellen. Die beiden Verfahren lassen sich miteinander zu einem kombinierten Verfahren vereinigen. In den folgenden Abschnitten wird zunächst die Akquise der verwendeten Trainings- und Testdaten beschrieben und anschließend die verschiedenen Indexierungsverfahren näher vorgestellt.

4.3.1 Trainings- und Testdaten

Sowohl für das Training des statistischen Verfahrens zur MeSH-Indexierung als auch für das spätere Testszenario werden Datensätze in Form von Artikelzusammenfassungen benötigt, die bereits manuell mit MeSH-Termen versehen sind. Diese wurden halbautomatisch von den PubMed-Webseiten sowie den Webseiten der Herausgeber extrahiert. Aus PubMed wurden hierbei die MeSH-Terme und die englischen Abstracts gewonnen, von den Originalseiten der Herausgeber wurden die Abstracts in den Originalsprachen der Artikel extrahiert. Die Anfrage in der PubMed-Suchmaske lautete wie folgt:

$< Language > [la]\ AND\ full\ text[sb]\ AND\ medline[sb]\ AND\ hasabstract, \quad mit$

$<Language>$	- Sprache der gewünschten Dokumente
$[la]$	- markiert das voranstehende Wort als Sprachbezeichner
full text	- gibt an, dass ein Link zur Originalseite existiert
$[sb]$	- Untermenge (Subset) von Medline
medline	- Dokumente sind bereits mit MeSH-Termen versehen
hasabstract	- Dokument besitzt ein Abstract

Aus diesen Anfragen wurden das englischsprachige Abstract mit Überschrift, die MeSH-Terme sowie der Originallink extrahiert. Mit Hilfe der Originallinks wurden anschließend die nicht englischsprachigen Abstracts samt Überschriften extrahiert. Tabelle 4.5 gibt einen Überblick darüber, wie viele Dokumente mit diesem Verfahren für die verschiedenen Sprachen erfasst wurden. Für das Englische stehen am meisten Dokumente zur Verfügung, da diese komplett aus PubMed extrahiert werden können, ohne einem Originallink folgen zu müssen. Da für die nichtenglischsprachigen Artikel nur wenige Zeitschriften in PubMed verlinkt sind, ist die

Anzahl der Artikel in anderen Sprachen deutlich niedriger. Zu beachten ist außerdem, dass sich dadurch eine Konzentration auf bestimmte Teilgebiete der Medizin nicht vermeiden lässt.

Tabelle 4.5: Überblick über die Anzahl der Trainings- und Testdaten in den verschiedenen Sprachen. Aus jeder Sprache wurden zufällig 500 Artikel als Testdaten aussortiert. Die restlichen Datensätze dienen dem statistischen Verfahren als Trainingskollektion.

Sprache	Artikel	Wörter
Englisch	35.600	7.394.920
Deutsch	4.008	657.147
Portugiesisch	1.362	243.540
Spanisch	1.498	373.148
Französisch	8.525	1.687.888
Total	50.993	10.356.643

In einem weiteren Schritt wurden aus allen Kollektionen zufällig 500 Artikel ausgesucht, die als Testkollektionen zur Verfügung stehen. Alle anderen Artikel dienen dem statistischen MeSH-Indexierungsverfahren als Trainingskollektionen. Listing 4.1 zeigt beispielhaft einen deutschen Artikel mit englischer MeSH-Indexierung. Außerdem sind dort bereits die vom heuristischen Verfahren vorgeschlagenen MeSH-Terme mit angegeben.

4.3.2 Heuristische MeSH-Indexierung - das MorphoMap-Programm

Für die regelbasierte Verschlagwortung von medizinischen Freitexten auf das MeSH-Vokabular mit Hilfe des MorphoSaurus-Systems wurde die Indexierungssoftware MorphoMap entwickelt. Durch die Verwendung von MorphoSaurus wird die mehrsprachige Dokumentenindexierung unterstützt, somit können Dokumente in allen von MorphoSaurus unterstützten Sprachen auf den englischen MeSH abgebildet werden. MorphoMap ist modular und generisch aufgebaut, prinzipiell wird daher die Verschlagwortung von biomedizinischen Dokumenten mit beliebigen Vokabularien ermöglicht. Die Ermittlung von potentiellen MeSH-Termen erfolgt dabei, wie bei regelbasierten Verfahren üblich, über einen Zeichenkettenabgleich. Es stim-

Listing 4.1: Beispiel eines MeSH indexierten Dokumentes aus der deutschen Testkollektion. Für das heuristische Verfahren sind die ersten 20 Ergebnisse dargestellt. Übereinstimmungen mit der manuellen Vergabe sind mit einem Stern (*) markiert.

```
<article>
 <data language="German">
  <header>Belegtes zur präoperativen Therapie des Rectumcarcinoms</header>
  <abstract>Bei der Prognose der Rectumcarcinome ist die chirurgische Therapie
    unter Einhaltung der onkologischen Radikalitätsprinzipien von
    entscheidender Bedeutung. Oberstes Therapieziel sollte die R0-Resektion
    sein. Neben der Monoblock-Entfernung des tumortragenden Enddarmabschnitts
    mit dem dazugehörenden Mesorectum werden ausreichende orale, aborale und
    laterale Sicherheitsabstände verlangt. Der Chirurg hat durch seine
    adaequate Operationstechnik mit der Senkung der Lokalrezidivrate einen
    wesentlichen Einfluss auf die Überlebenswahrscheinlichkeit seiner
    Patienten. Der Prognosefaktor ``Chirurg'' stellt damit, wie retrospektive
    und prospektive Untersuchungen bewiesen, neben dem Tumorstadium die
    bedeutende, statistisch unabhängige Rolle. Durch die präoperative
    Radiochemotherapie kann man sich eine Steigerung der Rate an
    kontinenzerhaltenden Operationen erhoffen. Endgültige Ergebnisse und
    eindeutige Vorteile gegenüber der postoperativen Therapiemodalität, die
    derzeit in einer Multizenterstudie geprüft stehen noch aus. Die prä- oder
    postoperative Radiochemotherapie mit 5-Fluorouracil im Stadium UICC II und
    III stellt, wie vom Konsensuspapier der Deutschen Krebsgesellschaft vom
    1.7.1999 empfohlen, im multimodalen Therapiekonzept für Patienten mit
    einem erhöhten Risiko an Lokalrezidiven und Fernmetastasen eine
    unverzichtbare Komponente dar.
  </abstract>
 </data>
 <mesh>
   <term id="D003131">Combined Modality Therapy</term>
   <term id="D004740">English Abstract</term>
   <term id="D006801">Human</term>
   <term id="D020360">Neoadjuvant Therapy</term>
   <term id="D012004">Rectal Neoplasms/surgery</term>
   <term id="D012004">Rectal Neoplasms/radiotherapy</term>
   <term id="D012004">Rectal Neoplasms/mortality</term>
   <term id="D012004">Rectal Neoplasms/drug therapy</term>
   <term id="D015996">Survival Rate</term>
 </mesh>
</article>

Ergebnisse des heuristischen Mapping-Verfahrens:

1. Rectal Neoplasms*                  11. Medical Oncology
2. Combined Modality Therapy*         12. Survival Rate*
3. Radiochemistry                     13. Mastectomy, Segmental
4. Radiosurgery                       14. Treatment Outcome
5. Prognosis                          15. Factor Analysis, Statistical
6. Carcinosarcoma                     16. Preoperative Care
7. Aftercare                          17. Radiotherapy
8. Colorectal Neoplasms               18. Neoplasm Recurrence, Local
9. Oncologic Nursing                  19. Statistics, Nonparametric
10. Radiation Oncology                20. Intraoperative Care
```

men jedoch nur in circa 30% der Fälle die relevanten Textpassagen wortwörtlich mit dem passenden MeSH-Term überein [Aronson2006]. In allen anderen Fällen gibt es zwischen den Textpassagen und den MeSH-Termen nur partielle (oder gar keine) Übereinstimmung, so dass die größte Herausforderung darin besteht, anhand dieser partiellen Übereinstimmung aus der großen Menge von MeSH-Kandidaten die relevanten Terme zu identifizieren und an eine vordere Stelle in der Ergebnisliste einzuordnen.

Verschiedene Trefferarten bei der heuristischen MeSH-Indexierung

In einer kleinen Studie, in der 301 Nominalphrasen auf das UMLS-Vokabular abgebildet wurden, untersuchte [Aronson2006] die möglichen Szenarien, die zu einem Treffer zwischen Freitext und UMLS-Vokabular führen. Die vier identifizierten Trefferarten lassen sich qualitativ auf die MeSH-Indexierung übertragen[3]:

- **Einfacher Treffer** (ca. 30%): Eine Passage im Freitext lässt sich exakt auf einen MeSH-Term abbilden. Beispielsweise tritt in einem deutschen Artikel der Begriff *"Retroperitonealraum"* auf, welchem der MeSH-Term *"Retroperitoneal Space [A01.047.025.750]"* zugeordnet wird.

- **Komplexer Treffer** (ca. 8%): Ein komplexer Treffer tritt auf, wenn ein Mehrwort-Ausdruck aus dem Freitext auf zwei oder mehrere MeSH-Ausdrücke abgebildet wird. Beispielsweise wird der Ausdruck *"Intensivmedizinischer Notfall"* auf die MeSH-Terme *"Intensive Care [E02.760.190.400]"* und *"Emergencies [C23.550.291.781]"* abgebildet.

- **Partieller Treffer** (ca. 38%): Oft können nur Teile von Textpassagen auf MeSH-Terme abgebildet werden. Beispielsweise ist für den Ausdruck *"Kristallthermographie"* der MeSH-Term *"Thermography [E01.370.350.800]"* vorgesehen. Auch sogenannte *Overmatches* fallen hierunter, bei denen Teile des MeSH-Terms an einem der beiden Enden nicht gefunden wird, so sind zum Beispiel die MeSH-Terme *"Diabetes Complications [C19.246.099]"* und *"Intraoperative Complications [C23.550.505]" Overmatches* der Textpassage *"Komplikationen"*.

- **Kein Treffer** (ca. 24%): Kein Teil der überprüften Textpassage stimmt mit dem MeSH-Term überein.

[3]Bei der Übertragung der UMLS-Experimente auf den MeSH ist mit einer Änderung der Prozentwerte der verschiedenen Trefferwerte zu rechnen

Identifizierung möglicher Kandidaten in MorphoMap

Für das Auffinden möglicher MeSH-Kandidaten wird zunächst der gesamte MeSH-Thesaurus in der Vorbereitungsphase der morpho-semantischen Normalisierung des MorphoSaurus-Systems unterzogen. Dabei werden sowohl die Main Headings als auch deren Synonyme berücksichtigt. Die normalisierten MeSH-Einträge werden so strukturiert, dass für alle Äquivalenzklassen die zugehörigen MeSH-Terme unmittelbar ersichtlich sind (siehe Abbildung 4.1, *A*, auf Seite 103). So wird zum Beispiel die Zuordnung der Äquivalenzklasse *"#morphine"* zu den MeSH-Termen *"Dependence, Morphine"; "Morphine"; "Receptors, Morphine"; "Morphine Derivates"* etc. bereits zu diesem Zeitpunkt angelegt.

In der Benutzerphase werden für alle Äquivalenzklassen aus dem zu indexierenden Dokument die möglichen MeSH-Einträge ermittelt und als potentielle Kandidaten eingestuft. In der anschließenden Beurteilung wird die Relevanz all dieser MeSH-Kandidaten ermittelt.

Relevanzbeurteilung möglicher Kandidaten

Nachdem für ein Dokument die MeSH-Kandidaten bestimmt sind, wird nun mit deren Relevanzbeurteilung begonnen, indem die Kontextumgebung der jeweiligen Äquivalenzklasse im Dokument näher betrachtet wird (siehe Abbildung 4.1, *B*). Die Wortreihenfolge spielt bei der Kontextbetrachtung keine Rolle, so dass bei der Suche nach *"Schmerzen im Abdomen"* (*#pain #adomen*) auch der MeSH-Term *"Abdominal Pain"* (*"#abdomen #pain"*) gefunden wird. MorphoMap berücksichtigt eine Fenstergröße von zehn Äquivalenzklassen als Kontext einer Äquivalenzklasse. Zur Beurteilung der Relevanz von MeSH-Kandidaten wird eine Reihe von Faktoren berechnet. Für diese Faktoren existiert eine Sortierhierarchie, anhand derer die MeSH-Kandidaten in eine Rangreihenfolge gebracht werden. Im Folgenden werden alle relevanten Faktoren näher beschrieben, wobei die Reihenfolge der Vorstellung ihrer Position in der Sortierhierarchie entspricht:

- **Longest Match Factor (LMF)**: Bisweilen tritt bei der Indexierung das Phänomen auf, dass Äquivalenzklassen aus dem zu indexierenden Dokument sowohl auf mehrere Einwort-MeSH-Terme als auch auf einen Mehrwort-MeSH-Term abbilden. Beispielsweise erscheinen die zu dem Wort *"Bauchschmerzen"* gehörenden Äquivalenzklassen *"#abdomen #pain"* unter anderem in den drei MeSH-Termen *"Abdomen", "Pain"* und *"Abdominal Pain"*. In diesen Fällen wird der MeSH-Term mit den meisten übereinstimmenden

Äquivalenzklassen (hier *"Abdominal Pain"*) den anderen MeSH-Termen vorgezogen, er erhält einen *LMF* von 1. Alle anderen MeSH-Terme, in denen nur einzelne Äquivalenzklassen auftreten (hier *"Abdomen"* und *"Pain"*), erhalten einen LMF von 0. Erscheint in einer Textpassage hingegen der Ausdruck *"Nierenschmerzen"* (*"#kidney #pain"*), so erhalten die MeSH-Terme *"Kidney"* und *"Pain"* jeweils einen LMF von 1, da der Begriff *"Kidney Pain"* im MeSH nicht definiert ist.

- **Inverse Document Frequency (IDF)**: Ein besonderes Problem bei der Dokumentenindexierung bereiten diejenigen MeSH-Terme, die nur auf eine einzige Äquivalenzklasse abbilden. Beispielsweise tritt in einem Dokument relativ häufig und unspezifisch das Wort *"Behandlung"* (#treat) auf. Dieses Wort wird unter anderem auf den MeSH-Term *"Therapeutics"* abgebildet, für welchen im MeSH das Synonym *"Treatment"* definiert ist. Allerdings wird für die wenigsten biomedizinischen Dokumente der MeSH-Term *"Therapeutics"* als Deskriptor verwendet. Es gibt Dutzende anderer Beispiele, in denen unspezifische Äquivalenzklassen auf MeSH-Terme mit zumeist einer Äquivalenzklasse abbilden, was die Performanz der Verschlagwortung deutlich senkt. Daher wurde für Äquivalenzklassen eine Gewichtung eingeführt, die die Relevanz dieser Klassen näher definiert. Diese Gewichtung wird als *Inverse Document Frequency (IDF)* bezeichnet und ist den Gewichtungsfaktoren des Information Retrievals entliehen. Der IDF-Wert ist der Kehrwert aus der Anzahl der MeSH-Headings, in denen eine Äquivalenzklasse vorkommt. Die Äquivalenzklasse *#treat* beispielsweise erscheint in 220 MeSH-Termen und hat damit einen IDF-Wert von $1/220 = 0,004545$. Damit ist diese Klasse sehr unspezifisch und die Wahrscheinlichkeit, dass sie ausgerechnet auf den MeSH-Term *Therapeutics* abbildet, ist relativ gering. Erst ein IDF-Wert von 0.5 hat sich in Experimenten als sinnvoller Grenzwert herausgestellt, ab dem MeSH-Einträge als relevant eingestuft werden können, die auf einzelne Äquivalenzklassen abbilden.

- **Phrase Factor**: Die Anzahl von Äquivalenzklassen in einem Dokument, die gleichzeitig in einem MeSH-Term auftreten, wird mit der Variable $Nr\_Mids\_In\_Term\_And\_Phrase$ belegt. Außerdem kann ein sogenanntes $Phrase\_Interval$ definiert werden, welches die Spanne zwischen der ersten und der letzten Äquivalenzklasse im Dokument angibt, die in einem einzigen MeSH-Term auftauchen. Der *Phrase Factor* (PF) ist schließlich definiert als der Quotient aus $Nr\_Mids\_In\_Term\_And\_Phrase$ und $Phrase\_Interval$. Als Beispiel wird der deutsche Ausdruck *"Die Leber*

des Patienten wurde transplantiert" betrachtet. Er wird zunächst in die Äquivalenzklassen *"#hepat,#patient,#transplant"* umgewandelt. Ein relevanter MeSH-Term hierzu ist *"Liver Transplantation"* (*"#hepat,#transplant"*). $Nr\_Mids\_In\_Term\_And\_Phrase$ ist in diesem Fall 2, das $Phrase\_Interval$ ist 3. Somit ergibt sich ein $Phrase\ Factor$ von 2/3.

- **Entry Factor:** Ein ähnlicher Faktor wie der *Phrase Factor* kann auch für die MeSH-Terme selbst definiert werden. Parallel zu dem $Phrase\_Interval$ wird ein $Term\_Interval$ definiert, welches die Gesamtanzahl der Äquivalenzklassen in einem MeSH-Term angibt. Der Quotient aus $Nr\_Mids\_In\_Term\_And\_Phrase$ und $Term\_Interval$ ergibt den $Entry\_Factor$. Als Beispiel dient der deutsche Ausdruck *"noduläre Hyperplasie"* ([*#nodul*, *#above*, *#plast*]), welcher auf den MeSH-Term "Focal Nodular Hyperplasia" (*"#focal, #nodul, #above,#plast"*) abgebildet werden kann. $Nr\_Mids\_In\_Term\_And\_Phrase$ ist 3, das $Term\_Interval$ ist 4, somit ergibt sich ein $Entry\_Factor$ von 3/4.

- **SortFactor** ist ein Mittelwert aus *Entry Factor* und *Phrase Factor*. Dabei wurde empirisch ermittelt, dass dem *Entry Factor* ein doppeltes Gewicht zu verleihen ist, um optimale Ergebnisse zu ermitteln. Somit wird der *Sort Factor* bestimmt aus $(2 * Entry\ Factor + Phrase\ Factor)/3$.

- **Title Factor:** Deskriptoren, die im Titel gefunden werden, erhalten eine höhere Wertigkeit als Deskriptoren aus dem Abstract.

- **Nr_Mids_In_Term_And_Phrase** wurde bereits für den $Phrase\_Factor$ und den $Term\_Factor$ verwendet und gibt die Anzahl der Äquivalenzklassen an, die gleichzeitig im Dokument als auch in einem MeSH-Term auftreten. Sind alle bisherigen in der Sortierhierarchie angewendeten Faktoren gleich, so wird an dieser Stelle der MeSH-Term mit der höheren Zahl von Äquivalenzklassen bevorzugt.

Zur Verdeutlichung wird die genannte Sortierreihenfolge noch einmal zusammengefasst: Für jeden MeSH-Kandidaten wird zunächst ermittelt, ob sein LMF-Wert gleich 1 ist. Ist dies der Fall, wird bestimmt, ob entweder zwei oder mehr Äquivalenzklassen eines MeSH-Terms gefunden wurden, oder aber ob der IDF-Wert des MeSH-Terms größer als 0.5 ist. Die Menge aller MeSH-Terme, auf die das zutrifft, wird anschließend anhand des *Sort-Factors* sortiert. Innerhalb dieser Hierarchie wird

noch einmal differenziert, ob der Term aus dem Titel oder aus dem Abstract gewonnen wurde. Sind mehrere MeSH-Terme noch immer an der gleichen Position, wird derjenige mit der größten Anzahl von Äquivalenzklassen nach vorne eingeordnet.

Zu beachten ist, dass diese Einordnung in eine Rangreihenfolge auf Ebene der MeSH-Synonyme erfolgt. Im letzten Schritt werden daher alle Synonyme, für die bereits weiter oben aufgelistete Synonyme existieren, aussortiert. Anschließend werden zu den Synonymen die jeweiligen MeSH-Headings ermittelt und ausgegeben (siehe Abbildung 4.1, *C*). In Listing 4.1 sind die ersten 20 Ergebnisse für einen Artikel aus der deutschen Testkollektion beispielhaft dargestellt.

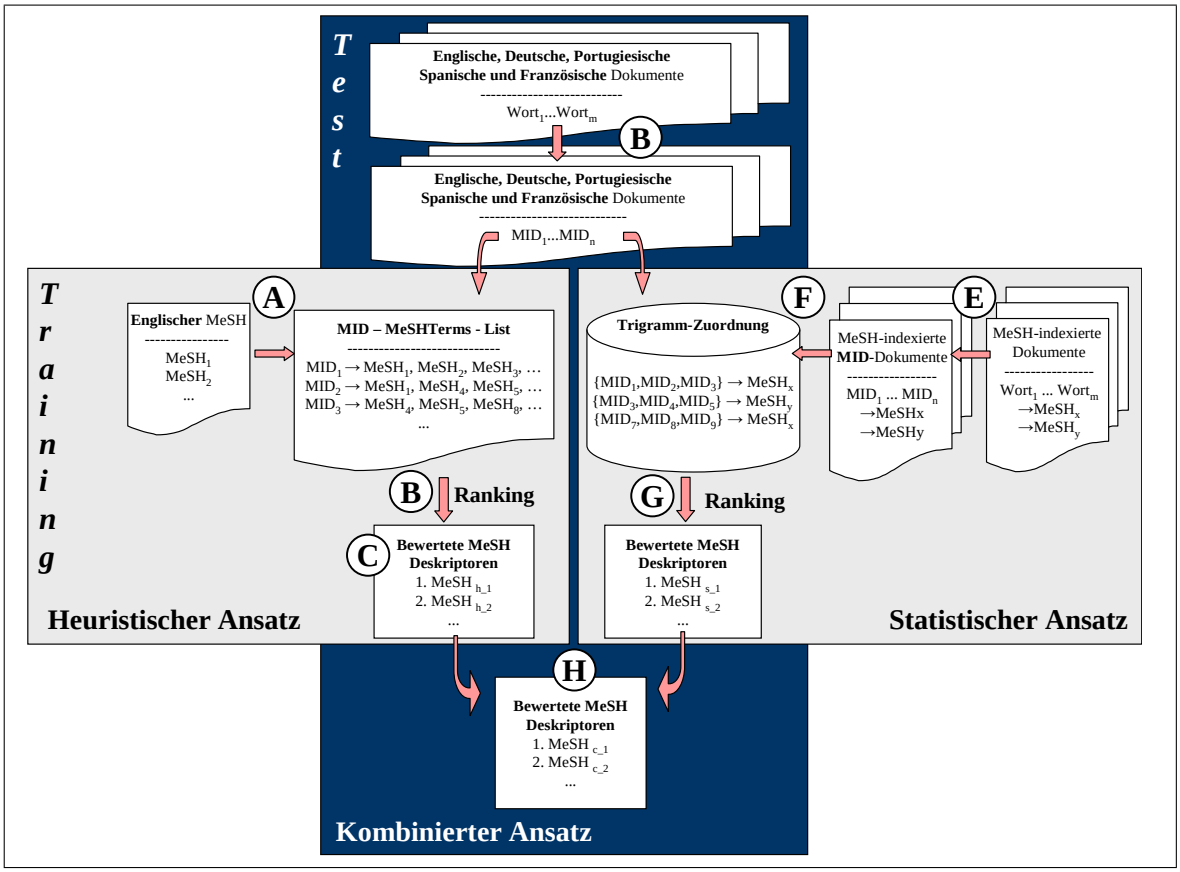

Abbildung 4.1: Übersicht über die Architektur des Indexierungssystems. Der heuristische Ansatz MorphoMap und der statistische Ansatz ermitteln unabhängig voneinander relevante MeSH-Deskriptoren. Diese werden anschließend zu einem gemeinsamen Ergebnis kombiniert.

4.3.3 Statistische MeSH-Indexierung

Als Ausgangspunkt für den statistischen Ansatz der MeSH-Indexierung dienen die in Abschnitt 4.3.1 beschriebenen Trainingsdaten, in denen den Dokumenten bereits manuell MeSH-Terme zugewiesen wurden (siehe Abbildung 4.1, *D*). Die Zahl der englischen Artikel ist mit ca. 35.000 Dokumenten deutlich am höchsten. In den anderen Sprachen sind zwischen 862 (für Portugiesisch) und 8.025 (für Französisch) Artikel verfügbar. Das statistische Verfahren ist ausführlich in [Markó2007] beschrieben.

Die Trainingsdaten werden zunächst in die MorphoSaurus-Interlingua überführt (siehe Abbildung 4.1, *E*). Anschließend werden Trigramme aus benachbarten Äquivalenzklassen gebildet und eine Zuordnung zwischen den Trigrammen und den manuell vergebenen MeSH-Termen erzeugt (siehe Abbildung 4.1, *F*). In allen Trainingsdokumenten werden diese Zuordnungen erzeugt und deren Häufigkeiten gezählt. Als Ergebnis entstehen Tripel bestehend aus MID-Trigramm, manuell vergebenem MeSH-Term und Häufigkeit F_{assoz} dieser Assoziation. Die Reihenfolge der Äquivalenzklassen in den Trigrammen spielt dabei keine Rolle, so dass auch Wortstellungs-Variationen in Texten mitberücksichtigt werden. Zusätzlich werden die *a-priori* Häufigkeiten $F_{apriori}$ von MID-Trigrammen ermittelt, die aussagen, wie häufig MID-Trigramme in der gesamten Trainingskollektion auftauchen.

In der Testphase wird für ein Testdokument auf Basis dieser Trainingsdaten die "wahrscheinlichsten" MeSH-Terme ermittelt, also diejenigen Terme, die bereits in den Trainingsdaten für ähnliche Dokumente manuell vergeben wurden. Das Testdokument wird zunächst in n Trigramme zerlegt. Anschließend werden für alle Trigramme diejenigen MeSH-Terme ermittelt, die bereits in den Trainingsdaten vergeben wurden. Für jeden dieser MeSH-Terme wird nun die Wahrscheinlichkeit P_{Rel} berechnet, mit der dieser Term für das Dokument relevant sein könnte. P_{Rel} dient dabei gleichzeitig als Gewichtungsfaktor für die Erstellung einer Rangreihenfolge zwischen den MeSH-Kandidaten (siehe Abbildung 4.1, *G*) und wird nach folgender Formel errechnet:

$$P_{Rel} = log \prod_{i=1}^{n} \begin{cases} \frac{F_{assoz}(i)}{F_{apriori}(i)} & \text{, falls definiert} \\ 1 & \text{, sonst} \end{cases}$$

Durch den Nenner $F_{apriori}$ in dieser Gleichung wird berücksichtigt, dass seltene Trigramme in der Regel eine größere inhaltliche Aussagekraft für ein Dokument besitzen als Trigramme, die in sehr vielen Dokumenten auftreten. Sind keine Informationen über die Assoziation von Trigrammen und MeSH-Termen aus den Trai-

ningsdaten ersichtlich, wird dieser Assoziation der Wert 1 zugeordnet und verhält sich in der Multiplikation der Gleichung neutral.

4.3.4 Kombiniertes Verfahren

Zusätzlich zu den zwei vorgestellten Ansätzen ist ein Ziel der Experimente herauszufinden, ob die Kombination beider Verfahren zu einer Verbesserung der Indexierungsergebnisse führt (siehe Abbildung 4.1, H). Hierzu wurde in einem empirischen Verfahren ermittelt, auf welche Weise die Verfahren am besten fusioniert werden können. Schließlich zeigte sich folgendes Kombinationsverfahren erfolgreich: MeSH-Einträge, die in beiden singulären Verfahren unter die ersten 30 Ergebnisse eingeordnet werden, werden an die vordersten Plätze der kombinatorischen Ergebnisliste gestellt. Aus den übrigen Ergebnissen werden abwechselnd zwei Einträge aus dem statistischen Verfahren und ein Eintrag aus dem heuristischen Verfahren entnommen und an die Ergebnisliste des kombinierten Verfahrens angehängt, bis schließlich 100 Ergebnisse für das kombinierte Verfahren zur Verfügung stehen.

4.4 Experimentelles Szenario

Für die Evaluation der Indexierungsverfahren wurden in den Sprachen Englisch (En), Deutsch (De), Portugiesisch (Pt), Spanisch (Es) und Französisch (Fr) 500 Artikel ausgewählt, die bereits manuell mit MeSH-Deskriptoren aus dem Jahr 2006 versehen wurden (siehe Abschnitt 4.3.1). Aus Gründen der Vergleichbarkeit wurden aus dem MeSH-Vokabular ähnlich wie in anderen Studien [Aronson et al.1999] die *Headings, Check Tags, Age Groups* und *Geographics* verwendet (siehe Abschnitt 4.2). Die Tests erweitern bereits früher durchgeführte Untersuchungen [Markó et al.2003, Markó et al.2004], in denen die Verfahren auf kleineren Test- und Trainingsdaten für die Sprachen Englisch, Deutsch und Portugiesisch durchgeführt wurden. Die MeSH-Deskriptoren der jeweils 500 Testartikel dienen als Goldstandard für diese Experimente, das heißt es wird überprüft, wie viele automatisch vergebene MeSH-Terme mit den manuell vergebenen MeSH-Termen übereinstimmen.

Die Verwendung der manuell annotierten MeSH-Deskriptoren als Gold-Standard ist nicht unkritisch. Wie in der Einleitung bereits beschrieben, beträgt die Konsistenz zwischen zwei menschlichen Indexierern (die sogenannte Interrater-Reliability) gerade einmal zwischen 48 und 61% [Funk & Reid1983] für MeSH-Headings. Solche Inkonsistenzen wirken sich negativ auf die Performanz der automatischen Verfahren aus, da davon ausgegangen werden muss, dass die automatischen Verfahren plausible

MeSH-Terme identifizieren, die bei der manuellen Indexierung nicht berücksichtigt wurden. Dennoch werden die manuell vergebenen MeSH-Terme auch in anderen Studien als Gold-Standard verwendet [Aronson et al.1999]), da geeignetere Gold-Standards nicht zur Verfügung stehen.

Die 500 Test-Artikel werden in jeder Sprache dem heuristischen (*Heur*), dem statistischen (*Stat*) und dem kombinierten (*Mixed*) Verfahren übergeben und dort mit englischen MeSH-Termen versehen. Bezeichnet werden die Testläufe nach den Namen der Verfahren, gefolgt von der Sprache der Testdokumente (*En, De, Pt, Es, Fr*). Damit ergeben sich folgende Namen:

- Heuristisches Verfahren: Heur-En, Heur-De, Heur-Pt, Heur-Es, Heur-Fr
- Statistisches Verfahren: Stat-En, Stat-De, Stat-Pt, Stat-Es, Stat-Fr
- Kombiniertes Verfahren: Mixed-En, Mixed-De, Mixed-Pt, Mixed-Es, Mixed-Fr

4.5 Ergebnisse der MeSH-Indexierung

In Tabelle 4.6 sind die Precision- und Recall-Werte für die Cut-Off-Punkte 5 (Top 5), 10 (Top 10) und 50 (Top 50) angegeben[4], Abbildung 4.3 stellt die Ergebnisse grafisch dar. Die Ergebnisse zeigen zunächst, dass der statistische Ansatz dem heuristischen Ansatz in allen Sprachen überlegen ist. Dies betrifft die Recall-Werte noch stärker als die Precision-Werte: Beim heuristischen Verfahren liegen die Recall-Werte bei einem Cut-Off von 50 nur zwischen 12% (Spanisch) und 28% (Englisch). Unter den ersten 50 Treffern des heuristischen Ansatzes wurde demnach höchstens jeder vierte MeSH-Eintrag auch bei der manuellen Indexierung vergeben. Im statistischen Verfahren liegen die Recall-Werte für Top 50 hingegen zwischen 47% (Englisch) und 55% (Deutsch). Damit ist jeder zweite manuell vergebene MeSH-Term unter den ersten 50 Treffern der statistischen Indexierung. Im kombinierten Verfahren sind die Recall-Werte im Vergleich zum statistischen Verfahren noch einmal um 1 bis 8 Prozentpunkte (Absolutwerte) erhöht.

Bezüglich der Precision-Werte sind die verschiedenen Verfahren etwas enger beieinander. Betrachtet man im heuristischen Verfahren die Precision-Werte bei einem Cut-Off von 5 in den verschiedenen Sprachen, so lässt sich darin gut die Abdeckung und Qualität der MorphoSaurus-Lexika ablesen. Wird die englisch-englische Indexierung mit einer Precision von 34% als Richtwert genommen, so kommt die deutsch-englische Indexierung mit einer Precision von 27% der einsprachigen Indexierung am

[4]Zu den Definitionen von Precision und Recall siehe 3.5

nächsten, gefolgt von der portugiesisch-englischen und französisch-englischen Indexierung mit 22%. Am schlechtesten schneidet die spanisch-englische Indexierung ab mit einer Precision von 11%.

Beim statistischen Ansatz spielt neben der Qualität der MorphoSaurus-Lexika vor allem die inhaltliche Kohärenz der verwendeten Trainings- und Testdaten eine entscheidende Rolle. Eine semantische Nähe innerhalb dieser Daten verbessert die Indexierungsperformanz eines statistischen Verfahrens. Da in den nicht-englischen Sprachen die Anzahl der in MEDLINE indexierten Zeitschriften weitaus geringer ist als im Englischen, sind damit auch die Dokumentenkollektionen aus einem kleineren Pool von Zeitschriften entnommen. Somit ist in den nicht-englischsprachigen Dokumenten eine semantische Nähe eher gegeben als für die englische Dokumentenkollektion. Dies könnte ein Grund dafür sein, dass die Ergebnisse der spanisch-englischen Indexierung im statistischen Verfahren mit 41% Precision bei Top 5 die besten Ergebnisse erzielen, noch vor der englisch-englischen Indexierung mit 38%, in deren Test- und Trainingsdaten eine Vielzahl heterogener Zeitschriften eingeflossen sind. Schlusslicht in diesem Verfahren bildet die portugiesisch-englische Indexierung mit 32% Precision. Das kombinierte Verfahren bietet wiederum einen absoluten Precision-Zugewinn von bis zu 7% (für Englisch).

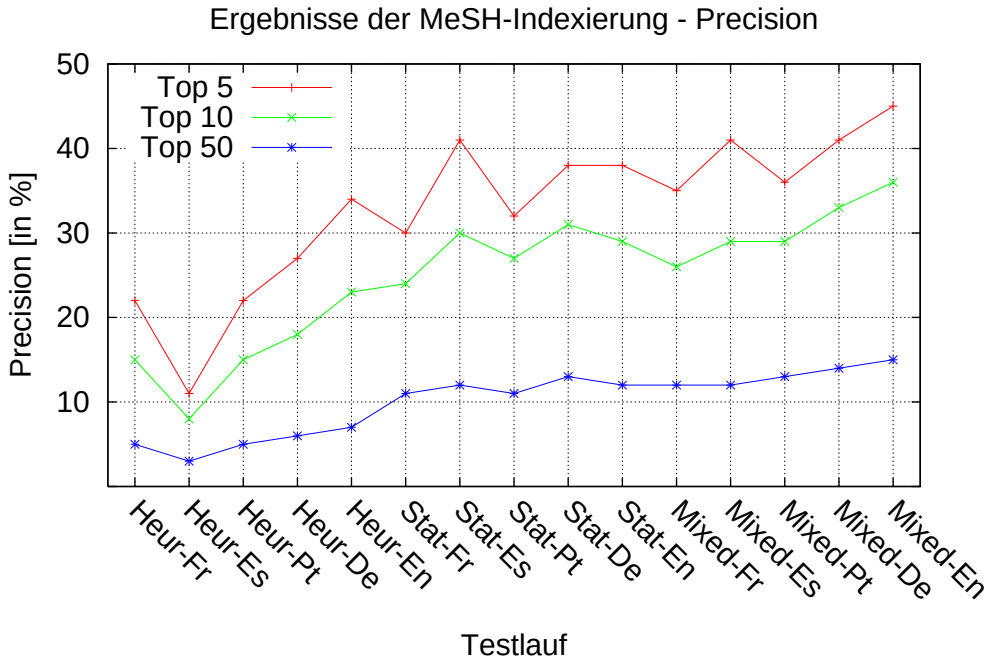

Abbildung 4.2: Übersicht über die Precision-Werte aller Testläufe für Top 5, Top 10 und Top 50

Tabelle 4.6: Precision/Recall Tabelle für alle Testläufe an den Cut-Off-Punkten Top 5, Top 10 und Top 50

Testlauf	**Top 5**		**Top 10**		**Top 50**	
	Prec	Rec	Prec	Rec	Prec	Rec
Heur-En	0,34	0,13	0,23	0,17	0,07	0,28
Stat-En	0,38	0,14	0,29	0,22	0,12	0,47
Mixed-En	**0,45**	**0,17**	**0,36**	**0,27**	**0,15**	**0,55**
Heur-De	0,27	0,11	0,18	0,15	0,06	0,25
Stat-De	0,38	0,16	0,31	0,26	0,13	0,55
Mixed-De	**0,41**	**0,17**	**0,33**	**0,28**	**0,14**	**0,59**
Heur-Pt	0,22	0,10	0,15	0,14	0,05	0,23
Stat-Pt	0,32	0,15	0,27	0,25	0,11	0,53
Mixed-Pt	**0,36**	**0,17**	**0,29**	**0,27**	**0,13**	**0,59**
Heur-Es	0,11	0,05	0,08	0,07	0,03	0,12
Stat-Es	**0,41**	**0,18**	**0,30**	**0,26**	**0,12**	0,53
Mixed-Es	**0,41**	**0,18**	0,29	0,25	**0,12**	**0,54**
Heur-Fr	0,22	0,10	0,15	0,14	0,05	0,23
Stat-Fr	0,30	0,14	0,24	0,22	0,11	0,50
Mixed-Fr	**0,35**	**0,16**	**0,26**	**0,24**	**0,12**	**0,54**

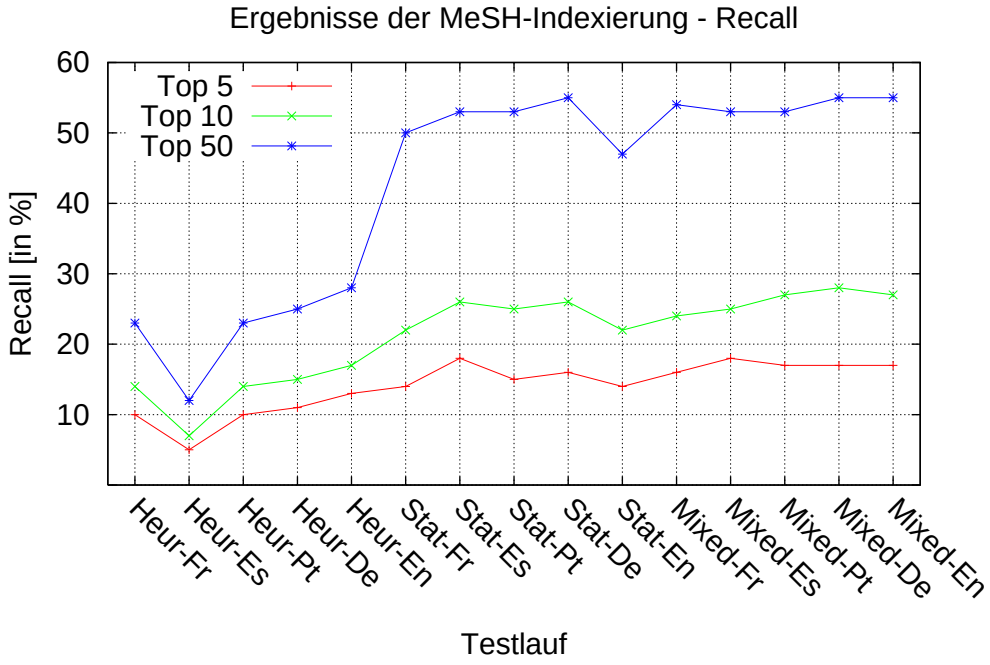

Abbildung 4.3: Übersicht über die Recall-Werte aller Testläufe für Top 5, Top 10 und Top 50

4.6 Diskussion der Ergebnisse und Ausblick

Zunächst soll untersucht werden, warum die Ergebnisse des regelbasierten Verfahrens im Vergleich zum statistischen Verfahren deutlich niedriger ausfallen. Bei qualitativer Beurteilung der heuristischen Indexierungsergebnisse zeigt sich, dass die Gründe hierfür vielfältiger Natur sind. Schaut man sich das Beispiel in Listing 4.1 genauer an, so fällt auf, dass die vorgeschlagenen MeSH-Terme *"3. Radiochemistry", "5. Prognosis", 6. "Aftercare" und 7. "Carcinosarcoma"* nicht grundsätzlich falsch sind. Aus verschiedenen Gründen wurden diese Terme jedoch nicht von den Indexierern ausgewählt, sei es, dass die Terme zu unspezifisch sind, sei es, dass sie den Indexierern nicht relevant genug erschienen oder sei es, dass die Indexierer diese Terme gar nicht in ihrem "aktiven Indexierungswortschatz" aufgenommen haben und diesen Term somit generell nicht vergeben[5]. Die Auswahl der relevanten Textpassagen sowie der MeSH-Terme ist also mit einem enormen intellektuellen Wissen verknüpft, welches sich durch jahrelange Erfahrung und durch Verinnerlichung der an der NLM aufgestellten Indexierungsmaßstäbe herausgebildet hat. Die regelbasierte Indexierungssoftware kann dieses intellektuelle Wissen nicht oder nur schwer reproduzieren. Beispielsweise behandelt es alle Wörter innerhalb eines Dokumentes

[5]In der Einleitung dieses Kapitels wurde bereits angesprochen, dass Indexierer bevorzugte Schlagwörter häufig, andere Schlagwörter jedoch wenig verwenden.

als gleich relevant, ein Maß für die Wichtigkeit eines Wortes innerhalb des Dokumentes existiert nicht. Dieses Problem wird durch den Einsatz des MorphoSaurus-Systems noch dadurch verstärkt, dass dieses System alle Wortformen wie Nomen und Verben gleich behandelt. Erscheint in einem Dokument beispielsweise die Phrase *"... wurden 100 Personen [...] behandelt"* (*"#patient #treat"*), so findet die Indexierungssoftware unter anderem den MeSH-Eintrag *"Treatment"* (*"#treat"*), obwohl dieser Deskriptor für das Dokument nur bedingt eine Rolle spielt. Durch den Einsatz von Wortartenerkennung und Chunking, könnte dieses Problem partiell behoben werden. So könnten beispielsweise relevante Nominalphrasen aus den Dokumenten extrahiert werden, bevor mit der MeSH-Indexierung begonnen wird. Dies ist unter den Annahmen sinnvoll, dass (1) die medizinisch relevanten Inhalte von freitextlichen klinischen Datensätzen in Nominalphrasen enthalten sind und dass (2) alle in den Nominalphrasen identifizierten Wörter auf das Zielvokabular abgebildet werden sollten [Cooper & Miller1998]. Allerdings entstehen dadurch auch neue Probleme, da zum einen die automatischen Verfahren zur Extraktion von Nominalphrasen selbst fehlerbehaftet sind, und da zum anderen relevante Informationen, welche nicht in den Nominalphrasen stecken, unberücksichtigt bleiben. So stellen [Won Kim2001] fest, dass ein Entfernen von anderen Wörtern als Stoppwörtern nicht zu einer Verbesserung der automatischen Indexierung führt.

Eine zweite Ursache für die moderaten Ergebnisse der heuristischen Indexierung liegt darin, dass für einige der manuell vergebenen MeSH-Deskriptoren keinerlei Wortbestandteile in den Dokumenten enthalten sind. In dem Artikel aus Beispiel 4.3.1 ist von *"präoperativer Radiochemotherapie"* die Rede. Wiederum ist es nur mit Hilfe intellektuellen Wissens möglich, aufgrund dieser Textpassage auf den MeSH-Term *"Neoadjuvant Therapy"* zu schließen. Die Indexierungssoftware hingegen erkennt lediglich *"Therapy"* als übereinstimmende Zeichenkette. Die Übereinstimmung von genügend großen Wortbestandteilen ist jedoch zentrale Voraussetzung für die Identifikation von MeSH-Termen durch die Indexierungssoftware.

Eine weitere Fehlerquelle entsteht durch fehlende oder semantisch zu grob gefasste Äquivalenzklassen des MorphoSaurus-Systems. So wird in den Subwort-Lexika bisher nicht zwischen einer *"therapeutischen Behandlung"* und der trivialsprachlichen Bedeutung von *"Behandlung"* unterschieden. So wird der MeSH-Eintrag *"Treatment"* auch gefunden, wenn in einem Dokument die Textpassage *"Dieser Artikel behandelt ..."* erscheint.

Positiv lässt sich herausstellen, dass das heuristische Verfahren zum einen unabhängig von Trainingsdaten anwendbar und sich somit auch auf andere Indexierungvokabularien übertragen lässt. Zum anderen trägt es in Kombination mit dem

statistischen Verfahren zu den Performanzgewinnen des kombinierten Verfahrens bei.

Das statistische Verfahren erzielt deutlich bessere Ergebnisse als der regelbasierte Ansatz. In diesem Verfahren ist kein Zeichenkettenabgleich zwischen Dokumenten und MeSH-Termen erforderlich, so dass auch semantisch ähnliche, aber textuell verschiedene Textpassagen und MeSH-Terme aufeinander abgebildet werden können. Da das statistische Verfahren die manuelle Indexierung "imitiert", in dem es häufige Textpassagen-Deskriptoren-Muster in den Trainingsdaten erkennt und diese in den Testdaten versucht wiederzufinden, ist es diesem Verfahren teilweise möglich, die intellektuelle Leistungen der Indexierer zu reproduzieren. Der große Nachteil statistischer Verfahren ist, dass eine ausreichend große Menge an Trainingsdaten zur Verfügung stehen muss, damit diese Verfahren anwendbar sind. Im Rahmen der MeSH-Indexierung ist dies gerade für die englische Sprache ausreichend gegeben, für andere Sprachen sind diese Daten jedoch nur eingeschränkt, mit teils erheblichem Aufwand und in geringem Umfang extrahierbar.

Wie bereits in der Einleitung dieses Kapitels angedeutet, ist ein zentrales Problem bei der hier durchgeführten Evaluation, dass kein universaler Referenzindex existiert [Lancaster1991]. Ein Dokument kann durch zwei unterschiedliche Mengen von Deskriptoren korrekt beschrieben werden, was die Inkonsistenzen zwischen den Indexierern, die in der Studie von [Funk & Reid1983] beschrieben sind, teilweise erklärt. [Névéol et al.2006b] suchen daher alternative Möglichkeiten zu den klassischen Precision/Recall basierten Evaluationsverfahren und messen die semantische Nähe zwischen automatischen und manuell vergebenen MeSH-Termen. Eine andere Möglichkeit der Evaluierung besteht darin, die automatisch indexierten Dokumente medizinischen Experten vorzulegen, die die Ergebnisse mit *relevant, partiell relevant, nicht relevant* beurteilen. Dabei besteht allerdings das Problem, das die Experten zwar relativ leicht beurteilen können, ob die gefundenen MeSH-Terme für ein Dokument relevant sind oder nicht, aber nur mit wesentlich mehr Aufwand ein Urteil darüber fällen können, welche relevanten Terme nicht gefunden wurden.

Als alternative Evaluation können die Ergebnisse der automatischen Indexierung auch in IR-Experimenten verwendet werden und untersucht werden, ob Dokumente mit automatisch vergebenen MeSH-Termen genauso gut oder gar besser als manuell indexierte Dokumente gefunden werden. Interessant sind hierbei die Ergebnisse mehrerer Studien, dass das Hinzufügen von Deskriptoren eines kontrollierten Vokabulars nicht immer zu einer Steigerung der Retrieval-Performanz führt und der Vorteil von MeSH-basierter Suche daher teils kontrovers diskutiert wird. [Hersh & Hickam1995] und [Yiming Yang1994] konnten zunächst keinen Vorteil ge-

genüber Freitextindexierung ermitteln. [Srinivasan1996a] führte ausführliche Metaanalysen verschiedener Ansätze durch und stellte fest, dass dem MeSH-Vokabular durchaus eine bedeutende Rolle im medizinischen IR zukommt. [Kim et al.2001] führten Retrieval-Experimente mit der von der NLM bereitgestellten Indexierungssoftware (siehe Abschnitt 4.7) durch und konnten zeigen, dass die automatisch basierte Indexierung zu gleich guten IR-Ergebnissen wie die manuell basierte Indexierung führt [Névéol et al.2006b]. [Jenuwine & Floyd2004] verglichen zwei Suchstrategien, eine basierend auf dem MeSH-Vokabular, eine andere basierend auf reiner Freitextsuche und kamen zu dem Ergebnis, dass durch MeSH-Termsuche weniger nicht relevante Dokumente gefunden werden (höhere Spezifität), dass durch Freitextsuche allerdings mehr relevante Dokumente gefunden werden (höherer Recall). Beide Verfahren fanden relevante Dokumente, die das andere Verfahren nicht als relevant kennzeichnete. Es scheint also, dass könnte die Retrieval-Performanz von Suchdiensten zumindest durch eine Kombination aus Freitextsuche und MeSH-Term basierter Suche deutlich erhöht werden kann. Ausführliche Experimente mit dem MorphoSaurus-System stehen hierzu noch aus.

Ein anderer Aspekt der automatischen Indexierung von Dokumenten zu IR-Zwecken ist, dass die leichten Performanzgewinne in aller Regel mit erheblichen Laufzeitverlusten einhergehen, da ein Großteil der Indexierungsverfahren Laufzeiten in der Größenordnung von Sekunden pro indexiertes Dokument aufweisen können. Bei Dokumentenkollektionen wie MEDLINE mit einer Größenordnung von mehreren Millionen Dokumenten ist dies durchaus nicht unproblematisch und erschwert derzeit noch den routinemäßiger Einsatz solcher Technologien.

4.7 Verwandte Arbeiten

Techniken zur Textkategorisierung wurden bereits intensiv erforscht und haben zu einer Vielzahl von Veröffentlichungen geführt (siehe [Yang1999] für eine Übersicht über verschiedene Ansätze). Zu den verschiedenen Techniken, die bei der automatischen Textklassifikation eingesetzt werden, zählen Naïve Bayes [McCallum & Nigam1998], k-Nearest Neighbor-Algorithmus [Yang1999], Supportvektormaschinen [Joachims1999], Boosting [Schapire & Singer2000] und Lernen von Klassifikations-Regeln [Apté et al.1994]. Die meisten dieser Systeme bilden Dokumente nur auf eine kleine Anzahl von Deskriptoren ab (bis zu einigen Hundert wie im Falle der *Reuters-21578-Kollektion* [Hayes & Weinstein1991]). Seltener wird auf eine Menge von über 100.000 Termen abgebildet, wie dies beim MeSH der Fall ist. Die Vergleichbarkeit derartiger Systeme ist nur bedingt sinnvoll, wie [Sebastiani2002]

betont. Viele verschiedene Parameter wie Größe und Struktur der Dokumentenkollektionen, Art der linguistischen Vorverarbeitung (Stemming, Wortartenerkennung, usw.) oder Wahl der Indexierungsmethode sowie der Parameter müssen berücksichtigt werden und erschweren eine objektive Vergleichbarkeit, auch im Hinblick auf fehlende Gold-Standards (siehe Diskussion in Abschnitt 4.6).

Die englischsprachige MeSH-Indexierung wird bereits seit den 1980ern thematisiert [Humphrey & Miller1987]. Die NLM führte mit ihrer Indexing Initiative die wohl ausführlichsten Untersuchungen zur englischen MeSH-Indexierung durch und entwickelte den Medical Text Indexer [Aronson et al.2000, Aronson et al.2004], der verschiedene statistische und regelbasierte Verfahren beinhaltet und den Indexierern Empfehlungen für die halbautomatische Verschlagwortung von MeSH-Deskriptoren geben soll. Das regelbasierte Verfahren mit dem Namen MetaMap indexiert Dokumente mit UMLS-Termen und führt anschließend Filterfunktionen durch, um die zugehörigen MeSH-Terme zu extrahieren. Im Gegensatz zu unserem Verfahren, bei dem die Ermittlung der MeSH-Kandidaten durch morpho-semantische Normalisierung des MorphoSaurus-Systems ermöglicht wird, ermittelt *MetaMap* die potentiellen Kandidaten durch Erzeugung von Ableitungsvarianten und unter Berücksichtigung von Synonymen, also durch Expansion der Wörter aus den Dokumenten. Für das Wort *"ocular"* werden zum Beispiel die Varianten *"oculus", "eye", "optic", "optical", "ophthalmic" etc.* erzeugt und mit einer Nähegewichtung bezüglich des Ursprungsworts versehen, die für die spätere Relevanzbeurteilung der MeSH-Terme verwendet wird. Ähnlich wie in unseren Experimenten wurden die verschiedenen Verfahren der NLM mit einem kleinen Testset von 200 Abstracts gegen einen manuell indexierten Gold-Standard verglichen. Die erzielten Ergebnisse der besten Kombinationsverfahren liegen bei Precision/Recall-Werten von 0,60/0,29 (Top5), 0,48/0,41 (Top10) und 0,20/0,61 (Top40). In neueren Experimenten mit 273 Medline Artikeln berichten [Aronson et al.2004] von Precision/Recall-Werten von 0,29/0,55 (Top25). In diesen Experimenten wurden außerdem Experten befragt, ob die 273 Artikel durch die MeSH-Terme ausreichend genau beschrieben wurden. 27% der Artikel wurden als ausreichend genau, 53% der Artikel partiell und 10% der Artikel als nicht ausreichend eingeschätzt. Der Recall lässt sich dabei noch um weitere 7% steigern, wenn Volltexte anstatt Abstracts verwendet werden [Gay et al.2005].

[Névéol et al.2005a] testeten drei französische MeSH-Indexierungswerkzeuge (CISMeF [Névéol et al.2006a], NOMINDEX [Pouliquen et al.2002] und HONMeSHMapper [Gaudinat & Boyer2002]) anhand von 82 Dokumenten aus dem CISMeF Katalog. Diese Dokumente wurden zuvor von fünf Indexierern manuell mit MeSH-Termen versehen, die als Gold-Standard für die Experimente dienen. Die

Precision/Recall-Werte für NOMINDEX liegen bei 0,13/0,9 (Top4), 0,13/0,23 (Top10) und 0,06/0,51 (Top50). Der HONMeSHMapper erreichte Werte von 0,32/0,26 (Top4), 0,21/0,37 (Top10) und 0,08/0,58(Top50). Schließlich erzielte CISMeF Werte von 0,31/0,22 (Top4), 0,21/0,37 (Top10) und 0,07/0,49 (Top50). Der HONMeSHMapper erreicht damit die besten Ergebnisse, wenn man das gewichtete Mittel aus Precision und Recall (sogenannter F-Wert) betrachtet.

In weiteren Experimenten testeten [Névéol et al.2005b] den CISMeF-Mapper, der in diesen Versuchen um eine statistische Komponente basierend auf dem k-Nearest Neighbor-Algorithmus erweitert wurde. Als Testkorpus dienten 52 Dokumente aus dem CISMeF Katalog, die in zwei Sprachen (französisch-englisch) vorliegen. Somit konnte dieses Verfahren mit dem englischen Medical Text Indexer der NLM verglichen werden. MTI erreichte in allen Szenarien bessere Ergebnisse als der CISMeF-Mapper, die besten Top-5 Precision/Recall-Werte lagen bei 0,41/0,46 für MTI und 0,31/0,33 für Französisch.

[Ruch et al.2003] kombinierten verschiedene Techniken zur MeSH-Indexierung, welche zum einen ein Verfahren zur Mustererkennung basierend auf regulären Ausdrücken beinhalten, zum anderen ein Vektorraummodell basierend auf den Gewichtungsverfahren $tf * idf$ und dem *Cosinus*-Ähnlichkeitskoeffizienten. Ihre Verfahren testeten sie in verschiedenen IR-Experimenten. Im direkten Vergleich mit dem MetaMap-Programm der NLM[6] erreichten sie etwas niedrigere Ergebnisse bezüglich der 11Pt-Avg-Precision.

Arbeiten von [Hersh & Donohoe1998] und [Zweigenbaum et al.2001] beschäftigen sich bei der MeSH- bzw. UMLS-Indexierung insbesondere mit der Verwendung morphologischer Analysetechniken wie Stemming oder Anfrageerweiterung durch morphologische Varianten und konnten den Nutzen einer derartigen morphologischen Vorverarbeitung zeigen. Für stark flektierende und agglutinierende Sprachen zeigten sich einfache Stemming-Algorithmen jedoch als nicht ausreichend [Hersh & Donohoe1998].

[6]MetaMap bildet in der eigentlichen Version auf UMLS-Deskriptoren ab. In diesen Experimenten wurde das von MetaMap verwendetet Vokabular jedoch auf die MeSH-Terme beschränkt.

Kapitel 5

Termübersetzung mit MorphoSaurus

5.1 Einleitung

Die Idee, dass Computersysteme von einer Sprache in eine andere übersetzen, ist beinahe so alt wie die Idee von Computersystemen selbst. Warren Weaver sprach bereits 1949 in dem als Weaver-Memorandum bekannt gewordenen Dokument "Translation" von der maschinellen Übersetzung. In jüngerer Zeit äußerten [Brown et al.1990] den Gedanken von der Möglichkeit von Systemen zur vollautomatischen Übersetzung. Hierbei ist die Kenntnis der in den verschiedenen Sprachen vorhandenen Wörter sowie ihrer interlingualen Verknüpfung elementare Voraussetzung zur automatischen Übersetzung. Eine zentrale Herausforderung ist die Erstellung genügend großer bilingualer Wortlisten, die Übersetzungen für die in einer Sprache vorhandenen Einzelwörter sowie für Mehrwort-Ausdrücke enthalten, sofern die Bedeutung der Mehrwort-Ausdrücke nicht von der Bedeutung der Einzelwörter abgeleitet werden kann.

Bei der automatischen Übersetzung können vereinfachend zwei Verfahren unterschieden werden, zum einen Wörterbuch basierte Übersetzungsverfahren, zum anderen Korpus basierte Übersetzungsverfahren. Wörterbücher werden in der Regel in aufwändiger Expertenarbeit erstellt. Sie haben den Vorteil, dass ein im Wörterbuch enthaltener Begriff in aller Regel richtig übersetzt ist. Nachteile dieses Verfahrens sind neben den Kosten zur Erstellung eines solchen Wörterbuchs der Umgang mit Wörtern, die nicht im Lexikon enthalten sind (sogenannte *Out-Of-Vocabulary-Terms, OOV*). Fehlende Begriffe stellen gerade in agglutinierenden Sprachen wie dem Deutschen, in denen durch das Aneinanderfügen einzelner Wörter fast beliebig

lange neue Wörter gebildet werden können, ein nicht unbeachtliches Problem dar.

Korpus-basierte Verfahren verwenden in der Regel parallele Korpora zur automatischen Übersetzung. Als Parallelkorpora werden Sammlungen übersetzter Texte bezeichnet. Sie zeichnen sich dadurch aus, dass genau ein übersetztes Wort für ein Wort des Ursprungstextes existiert, und dass Worthäufigkeiten und Positionen der Wortpaare in beiden Texten vergleichbar ist. Dementsprechend werden diese Worthäufigkeiten, Positionen oder auch Ähnlichkeiten in der Schreibweise dazu verwendet, Übersetzungspaare in den Parallelkorpora zu identifizieren [Rapp1999]. In Domänen, in denen keine Parallelkorpora vorhanden sind, werden verwandte Korpora oder sogar nicht verwandte Korpora eingesetzt. Verwandte Korpora unterscheiden sich in Worthäufigkeiten und Wortstellungen, aber haben dennoch ähnlichen Inhalt und Gebrauchsmuster[Fung & Yee1998]. Nicht verwandte Korpora weisen keine Gemeinsamkeiten auf.

Automatische Übersetzung ist ein wichtiges Ziel bei zahlreichen mehrsprachigen Anwendungen der natürlichen Sprachverarbeitung[Melamed2000]. Im Bereich der mehrsprachigen Dokumentenrecherche werden Übersetzungswerkzeuge dazu eingesetzt, Benutzeranfragen in eine andere Sprache zu übersetzen (*Query Translation*) [Daumke et al.2007b, Levow et al.2005, Oard & Diekema1998]. Im Bereich der mehrsprachigen Generierung natürlicher Sprache werden unter anderem multilinguale Terminologie-Server erstellt, die den Sprachtransfer in verschiedene Sprachen ermöglichen [Wagner et al.1995]. Schließlich kann ein Übersetzungswerkzeug im mehrsprachigen Kontext auch als Nachschlagewerk dienen beziehungsweise dazu verwendet werden, existierende Wörterbücher mit neuen Einträgen (halb-)automatisch zu ergänzen [Daumke et al.2005a]. Der Erfolg der Online-Wörterbücher der Leo GmbH[1] mit über 10 Mio. Anfragen pro Tag verdeutlicht die Relevanz derartiger Applikationen.

Wird der Begriff der Übersetzung nicht nur auf den Transfer zwischen verschiedenen Sprachen, sondern auch auf den Transfer zwischen verschiedenen Domänen innerhalb einer Sprache *"Slangs"* angewendet, finden sich gerade in der Medizin weitere wichtige Anwendungsmöglichkeiten. In den letzten Jahren hat die verbesserte Zugänglichkeit zu medizinischen Sachthemen im Internet zu Veränderungen im Arzt-Patienten-Verhältnis geführt. Patienten greifen zunehmend in den ärztlichen Entscheidungsprozess ein, um ihre medizinische Versorgung aktiv mitzugestalten. Dies setzt voraus, dass dem Laien hochwertige Informationen einfach zugänglich gemacht werden [CapGemini2005]. In diesem Zusammenhang kommt dem Transfer

[1] `http://dict.leo.org`, eingesehen im Februar 2007

von Experten- auf Laiensprache mit dem Ziel, ärztliche Informationen für den Laien verständlich zu machen, eine wichtige Aufgabe im Sinne einer patientenzentrierten Versorgung zu.

In diesem Kapitel wird ein neuartiger Ansatz zur automatischen Übersetzung von biomedizinischen Wörtern und Mehrwort-Termen vorgestellt. Die vom MorphoSaurus-System bereitgestellte Interlingua dient dabei als Brücke zwischen den verschiedenen Sprachen. Durch die Verwendung von MorphoSaurus wird die Größe der multilingualen Wortlisten, die zur automatischen Übersetzung verwendet werden, um ein Vielfaches reduziert und somit der Aufwand für Erstellung und Pflege der Ressourcen verringert. Die Qualität der Übersetzungen wird anhand von Testkollektionen aus dem UMLS [Umls2005b], einem Metathesaurus, der zahlreiche biomedizinische Terminologien umfasst, quantitativ evaluiert. Bis zu 88% korrekter oder ähnlicher Übersetzungen werden erreicht [Daumke & Markó2006].

5.2 Verfahren zur Termübersetzung

Die Übersetzung medizinischer Wörter oder Mehrwortausdrücke mit Hilfe von MorphoSaurus ist ein zweistufiges Verfahren: In der ersten Phase der Übersetzung wird eine Anfrage in die Interlingua von MorphoSaurus übersetzt (siehe Abbildung 5.1, E und F), während in der zweiten Phase der Transfer von der Interlingua in die Zielsprache erfolgt (Abbildung 5.1, F-J). Um diesen Transfer zu ermöglichen, werden in einer Vorbereitungsphase große Wortlisten in den unterstützten Zielsprachen erzeugt und in die MorphoSaurus-Interlingua überführt (Abbildung 5.1, A-D). Diese Wortlisten dienen bei der Übersetzung als *Übersetzungstabellen* von der Interlingua in die Zielsprache. Im Folgenden werden zunächst die Vorbereitungsphase und anschließend die Übersetzungsphase genauer beschrieben.

5.2.1 Erzeugung der Übersetzungstabellen

Zunächst wurden aus dem Internet große medizinische Textkorpora in verschiedenen Sprachen akquiriert (siehe Abbildung 5.1, Schritt A). Darin enthalten sind beispielsweise Abstracts aus medizinischen Zeitschriften aus MEDLINE[2] sowie Informationen aus verschiedenen Gesundheitsportalen wie beispielsweise NETDOCTOR [3] oder das Portal der Mayo-Kliniken[4]. Anschließend wurden die Dokumente normalisiert, was die Entfernung von HTML-Tags, die Umwandlung der diakritischen

[2] http://www.ncbi.nlm.nih.gov/entrez/, eingesehen im Februar 2007
[3] `http://www.netdoctor.co.uk/`, eingesehen im Februar 2007
[4] `http://www.mayoclinic.com/`, eingesehen im Februar 2007

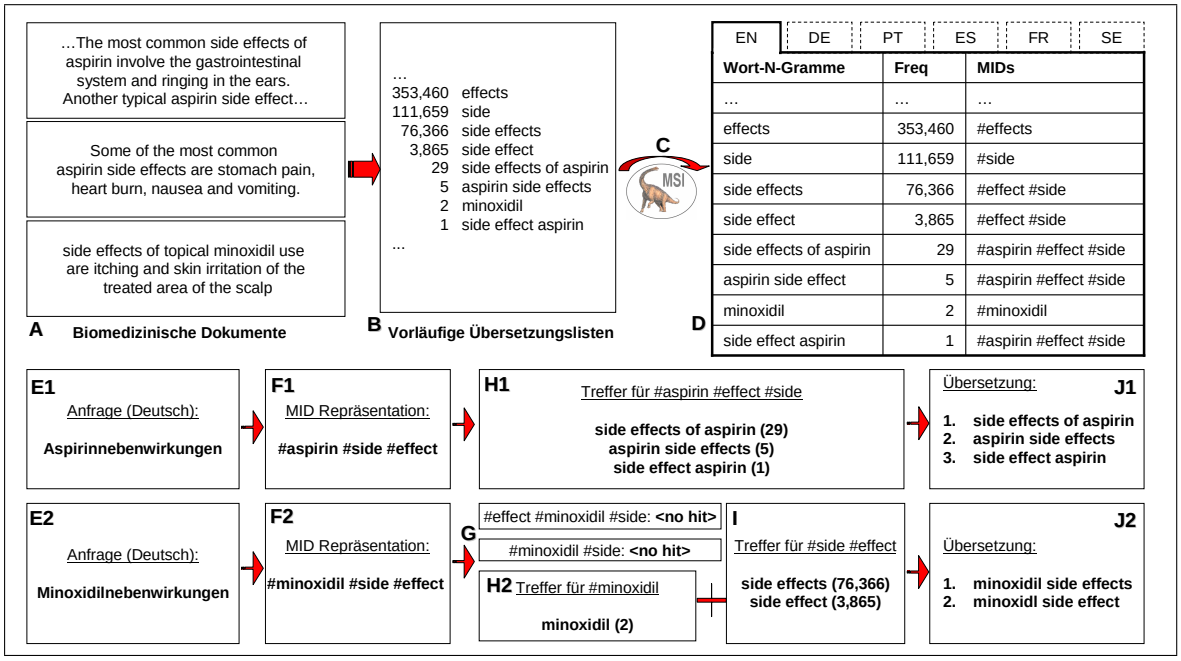

Abbildung 5.1: Überblick über die einzelnen Phasen der Termübersetzung. Die Schritte A-D werden in der Vorbereitungsphase durchgeführt. In E-J werden zwei ähnliche deutsche Anfragen (*Aspirinnebenwirkungen* bzw. *Minoxidilnebenwirkung*) beispielhaft ins Englische übersetzt.

Zeichen in 7-bit-ASCII-Code durch Anwendung sprachspezifischer Transformationsregeln (im Deutschen beispielsweise *'ß'* → *'ss'*, *'ä'* → *'ae'*, *'ö'* → *'oe'*, *'ü'* → *'ue'*) sowie die Umwandlung von Groß- in Kleinschreibung beinhaltet. In einem nächsten Schritt wurden die Korpora in Wort-N-Gramme zerlegt. Dabei wurde N auf Werte zwischen 1 und 3 beschränkt, so dass aus der Zerlegung der Korpora sprachspezifische Listen von Einzelwörtern, Wort-Bigrammen und Wort-Trigrammen resultieren. Anschließend wurden gleiche Wort-N-Gramme in diesen Wortlisten zusammen mit ihren Häufigkeiten zusammengefasst (Abbildung 5.1, B). Die Einträge dieser Liste werden nun der morpho-semantischen Normalisierung unterzogen (Abbildung 5.1, C) und in die sprachspezifischen Übersetzungstabellen übertragen (Abbildung 5.1, D), die folglich Tripletts bestehend aus Wort-N-Grammen, deren Häufigkeiten in den Korpora sowie ihrer MID-Repräsentation enthalten. Aufgrund der Häufigkeit von Subwort-Permutationen zwischen den einzelnen Sprachen wird die MID-Repräsentation zunächst alphabetisch sortiert. Damit wird erreicht, dass dem deutschen Wort *"Bluthochdruck"* und dem englische Ausdruck *"high blood pressure"* die gleiche MID-Repräsentation (*"#blood #high #pressure"*) zugeordnet wird. Tabelle 5.1 gibt einen Überblick über die Zahl der Wort-N-Gramme, die in den Sprachen Englisch (EN), Deutsch (DE), Portugiesisch (PT), Spanisch (ES),

Französisch (FR) und Schwedisch (SE) erzeugt wurden.

Tabelle 5.1: Anzahl der Wort-N-Gramme in den verschiedenen Sprachen.

Sprache	Einzelwörter	Bigramme	Trigramme
EN	528K	30.257K	97.673K
DE	467K	4.101K	5.530K
PT	138K	3.899K	7.058K
ES	125K	2.382K	3.746K
FR	85K	1.129K	1.796K
SE	47K	423K	782K

5.2.2 Der Prozess der Übersetzung

Ein zu übersetzender Ausdruck T_{orig} wird zusammen mit der Sprache dieses Ausdruckes sowie der gewünschten Zielsprache an das Übersetzungswerkzeug geschickt. Dort wird der Ausdruck in seine MID-Repräsentation T_{MID} überführt (Abbildung 5.1, F1/F2). Anschließend wird die MID-Repräsentation schrittweise mit der Übersetzungstabelle in der gewünschten Sprache verglichen. Dabei werden im ersten Schritt diejenigen MIDs der Anfrage verwendet, die den ersten drei Wörtern der Anfrage entsprechen[5]. Dies trägt der Tatsache Rechnung, dass in der Übersetzungstabelle maximal Wort-Tigramme enthalten sind. Man beachte, dass auch diese MIDs vor der Anfrage an die Übersetzungstabelle alphabetisch sortiert werden. Wird ein Eintrag in der Übersetzungstabelle gefunden, werden die zugehörigen Wörter ausgegeben und dieselben Schritte werden auf die nächsten MIDs angewendet. Im Fall, dass kein Eintrag in der Übersetzungstabelle existiert, wird die Anfrage an die Tabelle schrittweise um ein MID reduziert, bis die Anfrage erfolgreich ist. Im Falle, dass auch bei der schrittweisen Entfernung der MIDs kein Eintrag in der Übersetzungstabelle existiert, wird in der Hoffnung auf ein identisches Wort in den verschiedenen Sprachen, was zum Beispiel bei Eigennamen häufig der Fall ist, das Originalwort zurückgegeben. Anschließend wird mit den nächsten MIDs fortgefahren. Diese Prozedur wird solange wiederholt, bis schließlich alle MIDs mit den Einträgen in der Übersetzungstabelle verglichen wurden.

[5]Falls die Anfrage weniger als drei Wörter enthält, werden entsprechend weniger MIDs verwendet.

5.2.3 Übersetzungsalternativen

Die Übersetzungstabelle aus Abbildung 5.1 D verdeutlicht, dass MID-Sequenzen in den Übersetzungslisten in der Regel mehrfach auftauchen. Die MID-Sequenz *"#side #effect"* erscheint sowohl in *"side effects"* als auch in *"side effect"*. Dies kann dazu verwendet werden, verschiedene Übersetzungsalternativen zu erzeugen. Diese Alternativen werden erzeugt, indem eine Übersetzung einer entsprechenden MID-Sequenz mit den Übersetzungen der anderen MID-Sequenzen kombiniert wird. Das Resultat ist eine Liste möglicher Übersetzungsalternativen, deren Anzahl exponentiell zu der Anzahl der MID-Sequenzen wächst. Ein einfacher Ranking-Algorithmus berechnet auf Basis der Häufigkeiten der MID-Sequenzen schließlich einen Score, anhand dessen die Reihenfolge der Ausgabe der verschiedenen MID-Sequenzen bestimmt wird.

In dem Beispiel aus Abbildung 5.1 werden für den deutschen Ausdruck *"Aspirinnebenwirkungen"* drei mögliche Übersetzungen gefunden. In diesem Fall war bereits die erste MID-Anfrage *"#aspirin #effect #side"* (Abbildung 5.1, F) erfolgreich. Die Ergebnisse werden entsprechend ihrer Häufigkeiten sortiert (Abbildung 5.1, H1), so dass mit *"side effects of aspirin"* die vermeintlich am häufigsten verwendete englische Übersetzung an erster Stelle erscheint (Abbildung 5.1, J1). Im zweiten Beispiel, bei der Übersetzung von *"Minoxidilnebenwirkungen"*, sind mehrere Iterationen nötig, um den Term ins Englische zu übersetzen (Abbildung 5.1, G-I). Der erste Teil der Übersetzung wird für das MID *"#minoxidil"* gefunden (Abbildung 5.1, H2). Der restliche Teil (*"#side #effect"*) wird in der darauffolgenden Iteration übersetzt (Abbildung 5.1, I). Daraus resultieren die zwei möglichen Übersetzungen *"minoxidil side effects"* sowie *"minoxidil side effect"* (Abbildung 5.1, J2).

5.3 Experimentelles Szenario

Für die Evaluierung des Übersetzungswerkzeuges wurden Terme aus dem *U*nified Medical Language System [Umls2005b] verwendet, ein Metathesaurus, der zahlreiche unterschiedliche medizinische Terminologien beinhaltet und miteinander verlinkt. UMLS enthält circa eine Millionen Konzepte und über zwei Millionen Terme und Synonyme in verschiedenen Sprachen, die über einen gemeinsamen Identifier zu einem gemeinsamen Konzept zusammengefasst sind. Als Terme sind einzelne Wörter oder Mehrwort-Ausdrücke erlaubt.

Zunächst wurden aus dem UMLS alle Konzepte ausgewählt, die gleichzeitig in englischer, deutscher, portugiesischer, spanischer und französischer Sprache vorliegen. Aus dieser Liste wurden alle Abkürzungen und chemischen Ausdrücke entfernt,

da MorphoSaurus derzeit weder Abkürzungen noch chemische Ausdrücke effizient unterstützt[6]. Aus den restlichen UMLS-Konzepten wurden zufällig 200 ausgewählt. Die durchschnittliche Anzahl an Wörtern pro UMLS-Konzept betrug zwischen 1,35 für das Deutsche und 1,69 für das Spanische. Die ersten fünf Konzepte lauten im Englischen wie folgt: *"Sucrose", "Alouattinae", "Cestode Infections", "Clioquinol", "Antibiosis", "Hemangiosarcoma"*.

Die 200 Testdaten der nicht-englischen Sprachen wurden nun nach der in Abschnitt 5.2 beschriebenen Methode ins Englische übersetzt. Dabei wurde nur die erste Übersetzung ausgegeben, die übrigen Alternativen wurden verworfen. Diese Übersetzungen wurden von jeweils zwei medizinischen Experten, die gute Kenntnisse in der Ursprungssprache der Terme als auch in Englisch besitzen, auf Richtigkeit überprüft. Für jede der drei Übersetzungen konnte dabei folgende Auswahl getroffen werden:

- **Exakte Übersetzung** (COR): Dies wurde ausgewählt, wenn die Übersetzung exakt demjenigen Ausdruck entsprach, den der Experte als beste Übersetzung erwartete.

- **Ähnliche Übersetzung** (REL): Diese Alternative wurde ausgewählt, wenn die inhaltliche Bedeutung der Übersetzung korrekt war, jedoch die Wortreihenfolge oder die grammatische Form nicht dem Ursprungskonzept entsprach oder wenn beispielsweise Stoppwörter wie *"of"* fehlten. Außerdem wurde REL auch ausgewählt, wenn sich bei der Übersetzung ein Bedeutungswandel ergab, wenn das deutsche Wort *Antiobiose* (als *eine in einem Organismus stattfindende Reaktion*) zum Beispiel übersetzt wurde in das englische Wort *Antibiotics* (als *Medikament*).

- **Falsche Übersetzung** (ERR): Wenn die Übersetzung fehlerhaft war oder wenn das ursprüngliche UMLS-Konzept nicht übersetzt werden konnte, wurde diese Alternative ausgewählt.

Als Baseline (BASE) unserer Experimente wurde die Anzahl der identischen Cognates bestimmt, also die Anzahl der Konzepte, die im Englischen wie im Deutschen exakt gleich lauten.

[6]Der Umgang des MorphoSaurus-Systems wird derzeit erarbeitet, weitere Informationen hierzu sind in [Markó et al.2006] zu finden.

5.4 Ergebnisse der Termübersetzung

Tabelle 5.2 zeigt die Ergebnisse der Experimente, die in Abbildung 5.2 zusätzlich als Grafik dargestellt sind. Der Kappa-Wert ist ein Maß für die Übereinstimmung der medizinischen Experten, welche die Übersetzungen beurteilten. Die Werte zwischen 71,5% und 84,3% deuten dabei auf eine hohe Übereinstimmung hin.

Tabelle 5.2: Übersicht über den Anteil der korrekten, verwandten und falschen Übersetzungen (COR,REL,WRO) sowie der Baseline (BASE) (in Prozent). Übersetzt wurden 200 Ausdrücke aus dem Deutschen, Portugiesischen, Spanischen, Französischen und Schwedischen (DE, PT, ES, FR, SE) ins Englische. KAPPA ist ein Maß für die Übereinstimmung der Beurteiler.

Sprache	COR	REL	WRO	KAPPA	BASE
DE	65,7	13,2	21,1	74,5	34,5
PT	54,6	12,6	32,8	80,5	26,5
ES	50,0	12,7	37,3	75,3	27,5
FR	52,5	9,2	38,3	71,5	30,5
SE	51,4	12,3	36,3	84,3	25,5

Die Übersetzung vom Deutschen ins Englische schneidet mit 78,9% richtiger oder ähnlicher Übersetzungen am besten ab, gefolgt vom Portugiesischen mit 67,2%. Für die anderen Sprachen erreichte das Verfahren exakte oder ähnliche Übersetzungen in ca. 63% der Fälle. Die Baseline reicht von 25,5% für das Schwedische bis 34,5% für das Deutsche.

5.5 Diskussion der Ergebnisse und Ausblick

5.5.1 Fehleranalyse

Die Ergebnisse unserer Evaluation sind gerade für das Deutsche vielversprechend. Sie reflektieren in etwa den Stand der MorphoSaurus-Lexika in den einzelnen Sprachen. Da die Subwort-Lexika im Deutschen und im Englischen am vollständigsten sind und auch die bilinguale Verknüpfung bereits am weitesten fortgeschritten ist, werden bei der deutsch-englischen Übersetzung die besten Ergebnisse erreicht.

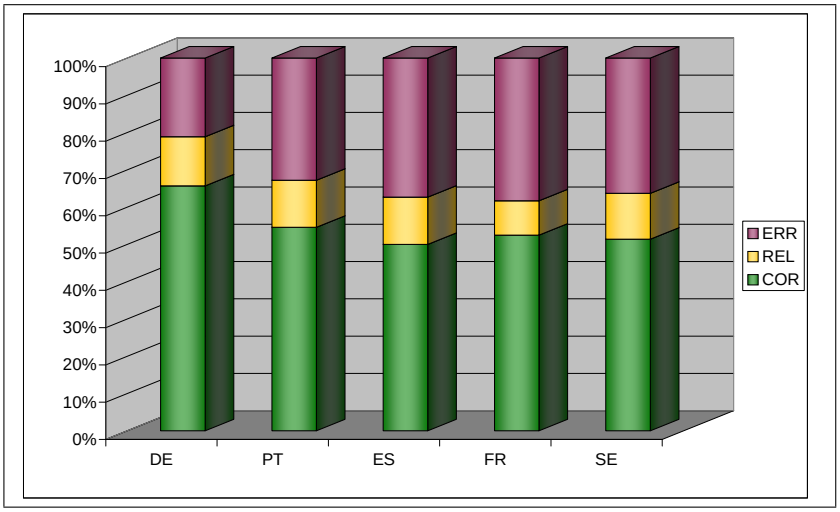

Abbildung 5.2: Grafische Übersicht über den Anteil der korrekten, verwandten und falschen Übersetzungen (COR,REL,WRO) sowie der Baseline (BASE) (in Prozent).

Wie die folgende Analyse zeigt, sind die meisten der falschen Übersetzungen durch fehlende Abdeckung in den Subwort-Lexika zu erklären. Dies ist insofern vielversprechend, als dass durch zusätzlichen lexikographischen Aufwand die Ergebnisse in allen Sprachen noch deutlich gesteigert werden können, und die Größe der lexikalischen Ressourcen dennoch um ein Vielfaches kleiner bleibt als bei Verfahren mit herkömmlichen Vollformlexika. Im Folgenden wird für die deutsch-englische Übersetzung eine ausführliche Fehleranalyse vorgestellt. Die Fehlertypen sind für die anderen Übersetzungspaare qualitativ die gleichen, lediglich die Prozentwerte können sich etwas ändern:

- **Unvollständige Abdeckung (89%)**: 32 Übersetzungsfehler entstanden durch fehlende Subwort-Einträge in den Subwort-Lexika. Die meisten davon waren Namen von Medikamenten oder Wirkstoffen wie zum Beispiel *"Barbital"*, die von MorphoSaurus derzeit noch nicht mehrsprachig unterstützt werden und daher falsch zerlegt werden. *"Barbital"* wurde von dem Übersetzungswerkzeug irrtümlich in *"Shaving Italy"* übersetzt.

- **Mehrdeutigkeit (11%)**: In vier Fällen wurde aufgrund von Mehrdeutigkeiten falsche Übersetzungen ausgegeben. Beispielsweise wurde das deutsche Wort *"Steuern"* zunächst in die MID-Repräsentation *"#control"* und *"#taxes"* überführt (im Sinne von *"Steuerung"* einerseits und *"Abgaben, Gebühren"* andererseits). Bei der anschließenden Übersetzung wurde das englische Wort *"control"* an erster Stelle ausgegeben, da dieses Wort die höchste Häufigkeit in der Übersetzungstabelle aufweist. In diesem Fall wäre *"taxes"* die richtige

Übersetzung gewesen.

Ein zweiter Fall von Mehrdeutigkeit tritt auf der Ebene von Mehrwort-Übersetzungen auf: Obwohl das Übersetzungswerkzeug die Wörter einzeln richtig übersetzt, ist der Gesamtausdruck dennoch inkorrekt. Der deutsche Mehrwort-Ausdruck *"Fötale Gefährdung"* (engl. *"fetal distress"*) wurde irrtümlich übersetzt in *"dangerous fetuses"* (*"Gefährliche Feten"*). Bei Betrachtung der Einzelwort-Übersetzung (*"Fötal"* → *"Fetuses"* bzw. *"Gefährdung"* → *"Dangerous"* wird deutlich, dass die einzelnen Wörter korrekt oder zumindest ähnlich übersetzt wurden.

Bei der Untersuchung der Abweichungen zwischen den einzelnen Beurteilenden (kappa-Wert) zeigte sich wiederholt, dass mit der Frage, ob ein Term korrekt oder ähnlich zu beurteilen ist, nicht immer einheitlich umgegangen wurde. Beispielsweise wurde der deutsche Ausdruck *"bösartige Neubildung der Niere"* mit der wörtlichen englischen Bedeutung *"malignant neoplasm of the kidney"* von dem Übersetzungswerkzeug in den alternativen, aber dennoch korrekten Ausdruck *kidney cancer* übersetzt. Einer der Experten beurteilte diese Übersetzung als korrekt, der andere als ähnlich.

5.5.2 Fazit und Ausblick

Vorgestellt wurde ein Verfahren zur automatischen Übersetzung von Wörtern und Mehrwort-Ausdrücken mit Hilfe der MorphoSaurus-Interlingua. Der besondere Vorteil dieses Verfahrens ist, dass mit Hilfe der Subwort-Lexika, die im Vergleich zu Vollformlexika relativ kleine Ressourcen darstellen, eine Vielzahl von Ausdrücken in andere Sprachen übersetzt werden kann. So ist es insbesondere möglich, Mehrwort-Ausdrücke, die in Lexika nicht auftauchen (OOV-Terms) automatisch korrekt zu übersetzen.

Der Erfolg dieses Verfahrens hängt entscheidend von der Abdeckung und der mehrsprachigen Verknüpfung der Subwort-Lexika ab. Da syntaktische Informationen wie Tempus, Modus oder Genus unberücksichtigt bleiben, eignet sich ein derartiges Verfahren vor allem in Bereichen, in denen diese syntaktischen Informationen keine Rolle spielen. So konnten bereits in einem Szenario zur multilingualen Dokumentenrecherche die Funktionalität dieses Verfahrens unter Beweis gestellt werden [Daumke et al.2005d]. Auf Basis dieses Übersetzungswerkzeuges wurde ein Web-Interface entwickelt (siehe Abbildung 5.3) [Daumke et al.2005b, Daumke et al.2005c], welches dieses Verfahren mit einem be-

kannten Internet-Suchdienst verknüpft[7]. Details zu diesem System können auch in [Daumke et al.2006, Daumke et al.2007b] entnommen werden.

Abbildung 5.3: MorphoSaurus Medical Web - Screenshot

Zukünftige Untersuchungen sollen die Integration syntaktischer Informationen berücksichtigen, um die Qualität der Übersetzungen weiter zu steigern. Ebenso könnten die verschiedenen Übersetzungsalternativen durch das Senden an den Internet-Suchdienst *Google* und anschließender Überprüfung, wie häufig eine vorgeschlagene Übersetzung tatsächlich im Internet erscheint, validiert werden. Die Hauptarbeit zur Verbesserung der Übersetzungsqualität liegt jedoch in der Ergänzung und Überarbeitung der lexikalischen Ressourcen von MorphoSaurus. Interessanterweise lässt sich gerade das hier vorgestellte Verfahren sehr gut für die Qualitätskontrolle und -verbesserung verwenden, wie der folgende Abschnitt verdeutlicht.

[7]Dieses Projekt wurde im Rahmen des Karl-Steinbuch-Stipendiums gefördert und wird daher an dieser Stelle nicht näher beschrieben.

5.5.3 Qualitätskontrolle der Subwort-Lexika

Ein positiver Nebeneffekt des vorgestellten Übersetzungswerkzeuges liegt darin, dass sich die lexikalischen Ressourcen von MorphoSaurus mit Hilfe der angebotenen Übersetzungsalternativen recht einfach und benutzerfreundlich evaluieren lassen und so eine Qualitätskontrolle der Lexika durchgeführt werden kann. Hierzu werden Wortlisten einer Sprache an das Übersetzungswerkzeug gesendet und verschiedene Übersetzungsalternativen in der gewünschten Zielsprache ausgegeben. Diese Zielsprache kann dieselbe wie die Ursprungssprache sein und zeigt dann eine Reihe von vorgeschlagenen Synonymen für die Eingabewörter. Diese können anschließend von Lexikographen revidiert werden. Im Folgenden werden einige fehlerhafte Synonymvorschläge für das Deutsche beispielhaft dargestellt:

Listing 5.1: Fehlerhafte Synonymvorschläge des Übersetzungstools

```
Herzwand:
herz wand    | herz wandler   | herz wandte    | herz waende | herz waenden |
herz gewandt | herz zugewandt | herzen wandler | herz wandten

Mittelschwere:
medikamente schwere  | media schwere  | arzneimittel schwere | mittelschwere

Venen:
venen    | vene  | venoesen | venoese | vena | venoeser | venedig |
venoeses | venae | venoesem | venous
```

Das Beispiel *"Herzwand"* zeigt, dass die Subwörter für *"wandler", "wandte", "gewandt", "zugewandt", "wandten"* noch nicht im Subwort-Lexikon enthalten sind. Beispiel 2 (*"Mittelschwere"*) macht die Ambiguität von Teilwörtern deutlich: *"Mittel"* wird in unserem Beispiel abgebildet auf *"Medikament"* und *"Arzneimittel"*, aber auch auf das lateinische Wort *"media"*. Erst weiter hinten in den Übersetzungsalternativen taucht die richtige Übersetzung *"mittelschwere"* auf. Der Übersetzungsfehler könnte vermieden werden, wenn *"mittelschwer"* als Subwort in unser Lexikon aufgenommen würde.

Im dritten Beispiel (*"Venen"*) wird als Synonym auch die Stadt *"Venedig"* vorgeschlagen. Grund dafür ist, dass von *"Venedig"* die Suffixe *"-d"* und *"-ig"* abgetrennt werden und schließlich der Stamm *"vene"* übrig bleibt. Der Fehler kann durch Aufnahme des Wortes *"Venedig"* als Subwort in unser Lexikon behoben werden.

5.6 Verwandte Arbeiten

Wie in der Einleitung dieses Abschnittes beschrieben, finden sich Ansätze zur Termübersetzung mit unterschiedlichem Fokus in verschiedenen Forschungsbereichen der natürlicher Sprachverarbeitung wieder, wie in der mehrsprachigen Dokumentenrecherche [Levow et al.2005, Oard1997], der Parallelanordnung von Korpora zur maschinellen Übersetzung [Resnik1999], bei der Wortdisambiguierung [Monz & Dorr2005, Kikui1998] oder auch bei der automatischen Lexikonaquise [Markó et al.2005b, Resnik & Melamed1997]. In allen diesen Bereichen ist die Übersetzung unbekannter Wörter, sogenannter *Out-Of-Vocabulary Terms*, eine besondere Herausforderung. Daher sind verschiedene Techniken zur Erweiterung bestehender Lexika und somit zur Minimierung der Out-Of-Vocabulary-Terms entwickelt worden. Als lexikalische Ressourcen werden entweder bilinguale Wörterbücher verwendet, die als sogenannte *Seed*-Lexika eingesetzt werden und mit Hilfe derer die bestehenden Ressourcen erweitert werden, oder es kommen mehrsprachige Parallel-Korpora [Melamed2000], verwandte Korpora [Fung1998] oder sogar nicht verwandte Korpora [Rapp1999] zum Einsatz. Bezüglich der textuellen Domänen, die in den Termübersetzungen zum Einsatz kommen, liegt der Fokus vor allem auf nicht wissenschaftlicher Literatur wie Zeitungstexten, weniger auf wissenschaftlichen Domänen wie der Medizin oder Jura [Oard2002]. Daher werden zunächst einige typische Ansätze allgemeinsprachlicher Übersetzungen diskutiert und daran anschließend einige biomedizinische Ansätze erwähnt.

Ein Problem bei der Mehrwort-Termübersetzung ist die Mehrdeutigkeit der einzelnen Wörter des Termes, die eine falsche Übersetzung von Teilen des Mehrwortausdruckes in die Zielsprache bewirken kann. In dem hier vorgestellten Verfahren wurden Fehler bei der Übersetzung von Mehrwort-Ausdrücken, in den mehrdeutige Wörter auftraten, dadurch minimiert, dass nicht nur Einzelwörter, sondern auch Wort-Bigramme und Wort-Trigramme in den Übersetzungstabellen verwendet wurden. [Kikui1998] setzt ebenfalls zweisprachige Wörterbücher sowie große Textkorpora in der entsprechenden Zielsprache ein, um Fehler durch Mehrdeutigkeiten bei der Übersetzung von Mehrwort-Begriffen zu minimieren. Zunächst werden in den Wörterbüchern die möglichen Übersetzungen der einzelnen Wörter ermittelt, anschließend wird mit Hilfe der Textkorpora in der Zielsprache bestimmt, welche Übersetzung die wahrscheinlichste ist. Dies wird anhand eines mehrdimensionalen Vektors ermittelt, der auf der Basis von Wort-Kookkurrenzen aus den Textkorpora erstellt wird. Die Methode wurde auf Termlisten angewendet, die Zeitungsartikel näher charakterisieren, und erreichte eine korrekte Wortdisambiguierung von 81%.

Die Verwendung eines Vektorraummodells sowie verschiedener Ähnlichkeitsmaße zur Bestimmung der wahrscheinlichsten Übersetzung ist auch der grundlegende Ansatz in der Arbeit von [Fung & Yee1998]. Sie konnten zeigen, dass die Assoziationen zwischen Wörtern und deren Kontextwörtern auch in nicht-parallelen Korpora erhalten bleiben. Das chinesische Wort für *Erkältung* tritt zum Beispiel häufig mit denselben chinesischen Kontextwörtern auf wie das englische Wort *"flu"* mit den entsprechenden englischen Kontextwörtern. Beide Wörter sind zum Beispiel häufig mit der jeweiligen Übersetzung des Wortes *"virus"* assoziiert. Wenn für diese Kontextwörter bereits Übersetzungen in Wörterbüchern bestehen, kann auf Basis zweisprachiger, nichtparalleler Korpora auch auf das Wort *"flu"* als Übersetzung des chinesischen Wortes für *"Erkältung"* geschlossen werden und die Wörterbücher dementsprechend erweitert werden. Dieser Ansatz ist sprachunabhängig und kann laut Autoren in verwandten Sprachen wie Englisch-Deutsch oder Englisch-Französisch angewendet werden.

Einen ganz ähnlichen Ansatz verwendet [Rapp1999] für die Suche nach OOV-Termen in nichtparallelen deutschen und englischen Texten. Ebenfalls auf Basis von bereits bekannten deutsch-englischen Übersetzungen werden mit Hilfe von Kookkurrenzen mögliche Übersetzungspaare aus nicht-parallelen Korpora extrahiert. Für ein unbekanntes deutsches Wort wird dazu der zugehöriger Kookkurrenzvektor mit Hilfe der bereits bekannten Übersetzungen in einen englischen Kookkurrenzvektor "übersetzt". Anschließend wird dieser übersetzte Kookkurrenzvektor mit allen englischen Kookkurrenz-Vektoren verglichen und eine Liste von möglichen englischen Kandidaten ausgegeben. In 72 von 100 Fällen wurden korrekte englische Übersetzungen an erster Stelle ausgegeben. Es sei angemerkt, dass in diesen Tests lediglich einzelne Wörter übersetzt werden können, jedoch keine Mehrwortausdrücke.

Ein anderer Ansatz zur Übersetzung unbekannter Terme wird von [Cheng et al.2004] vorgeschlagen. Sie verwenden bilinguale Ergebnislisten von Online-Anfragen an Internetsuchmaschinen als "Parallelkorpora". Dieses Verfahren nutzt die Tatsache aus, dass viele Termini der chinesischen Sprache zusätzlich in geklammerten Ausdrücken in Englisch angegeben werden. Die geeignete Übersetzung wählen sie anschließend mittels verschiedener Ähnlichkeitsmaße aus diesen Ergebnislisten aus. Ihr Verfahren verglichen sie mit einem herkömmlichen Wörterbuchbasierten Verfahren im CLIR-Kontext und konnten hierbei mit ihrem Verfahren bessere Ergebnisse bezüglich des MAP-Wertes (*Mean Average Precision*) erzielen. Auch [Zhang & Vines2004] nutzen in ihrem Ansatz die Tatsache aus, dass in chinesischen Online-Texten chinesische Texte häufig zusammen mit ihren entsprechenden englischen Übersetzungen auftreten. Für chinesische OOV-Terms wurden zunächst

Webanfragen in Chinesisch gestellt und die Ergebnisse anschließend auf vorhandene englische Wörter in der Nähe der chinesischen OOV-Terms untersucht. Anschließend wurde mit statistischen Analysen die wahrscheinlichste Übersetzung extrahiert. In CLIR-Experimenten konnten 22 von 25 OOV-Terms übersetzt und die Retrievalergebnisse signifikant gesteigert werden.

In der biomedizinischen Domäne basieren einige Ansätze zur Termübersetzung darauf, dass biomedizinische Begriffe in unterschiedlichen Sprachen häufig lateinischen oder griechischen Ursprungs sind und somit gleiche Wurzeln haben. [Schulz et al.2004] beschreiben ein Verfahren, bei dem Begriffe basierend auf einfachen Regeln zur Zeichenkettenumwandlung vom Portugiesischen ins Spanische übersetzt werden. Beispiele für diese Umwandlungen sind $qua \rightarrow cua$, $eia \rightarrow ena$ oder $lh \rightarrow j$. Für jeden der durch diese Unformung gewonnenen Kandidaten wird überprüft, ob er in Textkorpora der Zielsprache auftritt. In Experimenten wurde eine Genauigkeit dieses Verfahrens von 89,4% erreicht.

[Claveau & Zweigenbaum2005] verwenden einen Algorithmus, der Umformungsregeln aus französisch-englischen Termpaaren (Trainingsset) generiert. Dabei werden nur Termpaare verwendet, die morphologisch verwandt sind. Die Umformungsregeln werden iterativ wiederholt, wobei das Trainingsset jedesmal leicht verändert wird. So werden eine Vielzahl von Umformungsregeln erhalten. Anschließend wird das Testset anhand dieser Umformungsregeln "übersetzt". Verschiedene Umformungsregeln können dabei das gleiche Ergebnis liefern. Je häufiger ein Term gleich übersetzt wird, desto häufiger ist die Wahrscheinlichkeit der korrekten Übersetzung. In Experimenten wird eine Genauigkeit von 85% korrekt übersetzten Termen erzielt, wobei in den Tests nur einfache Wörter, jedoch keine Mehrwort-Ausdrücke verwendet werden.

[Chiao & Zweigenbaum2002] übertrugen den Ansatz von [Rapp1999], der mit Hilfe von ähnlichen, nichtparallelen Korpora und initialen Seed-Lexika semantisch ähnliche Terme findet, auf die biomedizinische Domäne. Als Kontextvektoren, die für jedes Wort in den Korpora berechnet werden, nutzen sie Ähnlichkeitsmaße wie *Häufigkeit*, $tf * idf$ und *log likelihood*. Getestet wurde ihr Verfahren auf 95 französischen Wörtern, die ins Englische übersetzt wurden. In 20% der Testwörter wurde die richtige Übersetzung an erster Stelle ausgegeben, in 50% der Fälle war die richtige Übersetzung unter den ersten zehn Kandidaten.

Kapitel 6

Einbettung von MorphoSaurus in das Informationssystem der Hautklinik Freiburg

6.1 Einleitung

Das medizinische Arbeitsumfeld ist gekennzeichnet durch papiergebundene und digitale Informationsüberflutung sowie durch eine qualitative Informationsunterversorgung. Im Rahmen des Projektes "Bedarfsgerechte Unterstützung von Ärzten an ihrem Arbeitsplatz über informationslogistische Anwendungen" wurden Ärzte über ihren Informationsbedarf und ihr Verhalten bei der Informationsbeschaffung befragt [Koch & Kaltenborn2005b]. Laut dieser Studie schätzen sie ihren Bedarf sowohl an patientenbezogenen als auch an nichtpatientenbezogenen Informationen als sehr hoch ein. 40% der Ärzte nutzen einmal bis mehrmals täglich das Internet als Wissensquellen bei der Informationssuche. 36% der Ärzte verbringen mindestens drei Stunden in der Woche, weitere 40% mindestens sechs Stunden pro Woche für Recherche und Durchsicht von Informationen. Der hohe Zeitaufwand für die Informationssuche, fehlende Informationen und der Aufwand für die papierbasierte Dokumentation führen zu einer deutlichen Unzufriedenheit der Ärzte sowie zu längeren Wartezeiten und potentiellen Fehl- bzw. Mehrfachuntersuchungen. Langfristige Auswirkungen der fehlenden Informationsverbreitung sind zum einen die Durchführung von Behandlungen, die sich in der aktuellen Medizin als ineffizient herausgestellt haben, oder aber die fehlende Durchführung von aktuellen Therapieempfehlungen [Hersh2002].

Die bisherigen Lösungen zur Informationssuche schätzen die Ärzte als zu kom-

plex und schwer bedienbar ein. Daher liegt die Abbruchquote von Recherchen im Durchschnitt bei 30%. Als häufigste Probleme bei der Recherche werden angegeben:

- das unübersichtliche Informationsangebot (79%)
- die häufig ungenauen Suchergebnisse (74%)
- die unsichere Qualität der Ergebnisse (70%)
- die lange Beschaffungsdauer (70%)

60% der Ärzte wünschen sich eine einfache und intuitive Bedienung von Datenbanken, wovon sie sich eine signifikante Senkung der Abbruchquote versprechen. Es besteht ein großes Interesse, auf Befunde, Arztbriefe, Röntgenbilder und andere im Krankenhaus erhobene Daten einrichtungsübergreifend zuzugreifen. Zu den Verbesserungsvorschlägen zählen Ärzte außerdem die Anbindungsmöglichkeiten ihrer Krankenhausinformationssysteme an externe Informationssysteme wie medizinische Datenbanken, die relevante wissenschaftliche Informationen in Volltextdarstellung enthalten. Dabei werden integrierte Lösungen in Form einer Metasuche gewünscht, die eine Abfrage über alle verschiedenen Datenbestände vornimmt ("Google für Ärzte"). Inhaltlich besteht ein großes Interesse an Leitlinien, fachmedizinischen und pharmazeutischen Informationen. Das Internet wird als Zugangsmedium zu solchen Informationen akzeptiert und häufig benutzt. Nicht zuletzt erhoffen sich die Ärzte, durch den Einsatz innovativer Technologien zur Informationssuche und -beschaffung mehr Zeit für Patienten zu gewinnen und die Qualität der Behandlung zu verbessern.

Die derzeitigen Informationssysteme sind für derartige Anforderungen jedoch nur bedingt ausgelegt, wenn man sich ihre typischen strukturellen Charakteristika verdeutlicht. "Ein *Krankenhausinformationssystem (KIS)* ist das Teilsystem eines Krankenhauses, das alle informationsverarbeitenden (und speichernden) Prozesse und die an ihnen beteiligten menschlichen und maschinellen Handlungsträger in ihrer informationsverarbeitenden Rolle umfasst. Das KIS dient dazu, die Mitarbeiter des Krankenhauses bei der Erledigung der Aufgaben des Krankenhauses zu unterstützen"[Winter et al.2002]. Die *elektronische Krankenakte* stellt berufsübergreifend alle einen Patienten betreffenden Informationen zusammen. Sie stellt damit den inhaltlichen Kern eines Krankenhausinformationssystems dar. Nach dieser Definition steht der patientenzentrierte, horizontale Zugang auf Informationen im Vordergrund eines KIS (siehe Abbildung 6.1). Die Behandlung eines Patienten kann von seiner Aufnahmeuntersuchung, über Labor-, Radiologie- und anderen Befunden bis hin zu seiner Entlassung nachvollzogen werden. [Dadam et al.2000] sehen

darin einen zunehmenden Wandel bestehender KI-Systeme hin zu einer Workflow-orientierten, direkten Unterstützung von Geschäftsprozessen.

Zusätzliche Aspekte der elektronischen Patientenakte werden durch die Healthcare Information and Management System Society (HIMMS)[1] ergänzt: "Die Patientenakte ist eine sichere, patientenzentrierte Echtzeit-Informationsquelle für Kliniker. [...] Neben der direkten Patientenversorgung unterstützt sie auch die Verwendung der Daten für andere Zwecke. Hierzu gehört das Rechnungswesen, das Qualitätsmanagement, die Auswertung klinischer Tests, die Ressourcenplanung sowie Beobachtungen und Auswertungen von Volkskrankheiten."

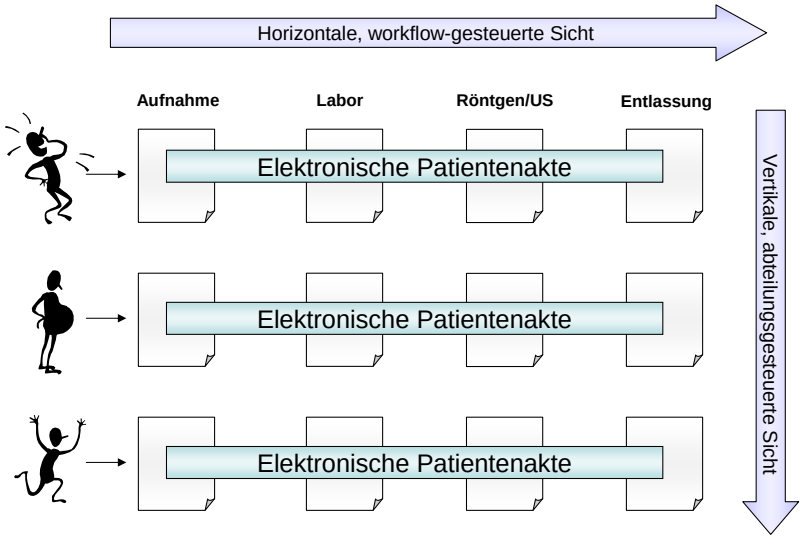

Abbildung 6.1: Verschiedene Sichtweisen auf ein Krankenhausinformationssystem

Die Definition der HIMSS legt neben der horizontalen Sichtweise auf das KIS besonderen Wert auf Querschnitts-Informationen im Sinne einer vertikalen, bereichsorientierten Sichtweise (Aufnahmeberichte, Laborbefunde, Radiologiebefunde, Verlaufsdokumentation etc.). Diese Querschnittsinformationen können strukturierte Einträge wie Diagnosen- oder Prozedurenschlüssel, Labor- und Mikrobiologiebefunde oder Abrechnungsübersichten enthalten. Trotz einiger vielversprechender Ansätze zur Standardisierung und Strukturierung klinischer Informationen wie die von HL7 erarbeitete, auf XML basierende Dokumentenarchitektur zur Übermittlung klinischer Inhalte (*Clinical Document Architecture*) [Dolin et al.2006, Mueller et al.2003], spielen aber auch unstrukturierte Daten im ärztlichen Alltag nach wie vor eine wichtige Rolle. Dazu zählen Entlassbriefe, Notizen über Anamneseerhebung, Befundberichte diverser Untersuchungen und andere Freitext-

[1] HIMSS Electronic Health Record, Definitional Model, Version 1.0: `http://www.himss.org/content/files/EHRAttributes.pdf`, eingesehen im Februar 2007

Informationen - je mehr Information im KIS enthalten sind, desto interessanter wird die vertikale, bereichsorientierte Sicht in das KIS [Müller et al.2007, Daumke et al.2007a].

Gerade auch im Hinblick auf Qualitätsmanagement und Kostencontrolling ergeben sich typische Fragen, die ein KIS beantworten sollte:

- "Welche Patienten wurden bisher mit der Krankheit X behandelt?"
- "Trat bei allen Patienten ein einheitliches Muster von Symptomen auf?"
- "Gab es Abweichungen in der Therapie bei diesen Patienten?"
- "Wie war der durchschnittliche Behandlungserfolg dieser Behandlungen?"
- "Welche Nebenwirkungen traten bei den Behandlungen auf?"

Durch die Einbettung des MorphoSaurus-Systems in das Informationssystem der Hautklinik Freiburg wird sowohl der Behandlungsverlauf einzelner Patienten (horizontale Sichtweise) als auch der Blick auf einzelne Bereiche (vertikale Sichtweise) durch eine freitextbasierte Suche zugänglich gemacht. Dabei wird insbesondere der Forderung der Ärzte nach einer einfachen bereichsübergreifenden Suche ("Google für Ärzte") Rechnung getragen. Die Ärzte sollen die Möglichkeit erhalten, über ein einziges Interface Behandlungsinformationen von Patienten wie Entlassbriefe, Labor- und Radiologiebefunde, aber auch wissenschaftliche Forschungsergebnisse sowie einschlägige Standardwerke durchsuchen zu können. Die morpho-semantische Indexierung des MorphoSaurus-Systems (siehe Kapitel 2.6) soll dabei die Trefferqualität der Suche und somit die Akzeptanz des Systems bei den Ärzten erhöhen. Im Folgenden wird die Einführung des MorphoSaurus-Systems in das KIS sowie die in der Hautklinik durchgeführte Evaluation dieses Systems näher beschrieben.

6.2 Integration von MorphoSaurus in das Informationssystem

6.2.1 Klinische Datensätze

Aus dem klinikeigenen Informationssystem, in welchem Daten von insgesamt 1,3 Mio. Patienten gespeichert werden, wurden alle 30.000 für die Hautklinik relevanten klinischen Dokumente seit dem Jahr 2000 extrahiert. Diese Dokumente liegen in Rich Text Format (RTF) vor und beinhalten Entlassbriefe, chirurgische OP-Berichte, immundermatologische Befundungen und andere freitextliche Dokumente. Durch ein

Pull-Verfahren werden neue Dokumente, die in der täglichen Routine der Hautklinik erstellt werden, wöchentlich vom KIS in das IR-Modul übertragen.

Gerade in der Dermatologie spielt die fotografische Befundung von Hautveränderungen beispielsweise zur Verlaufskontrolle eine wichtige Rolle. In der Hautklinik Freiburg liegen diese Bilder für den Zeitraum bis Juni 2006 in einer selbst entwickelten Access-Datenbank vor. Diese Datenbank umfasst einen Bestand von 90.000 Bildern, welche dem IR-System uneingeschränkt zur Verfügung stehen. Ab Juli 2006 wurde diese Datenbank durch ein proprietäres Archiv- und Dokumenten-Managementsystem [2] abgelöst. Die darin enthaltenen Bilder sollen in den kommenden Wochen ebenfalls in das IR-System integriert werden.

6.2.2 Künftige Erweiterungen

Als Proof-of-Concept wurden als wissenschaftlichen Literaturdatenbanken zunächst SOMED[3] und HECLINET[4] hinterlegt. SOMED ist eine deutsch- und englischsprachige Literaturdatenbank, auf den Gebieten Sozialmedizin und Public Health. HEalth Care Literature Information NETwork (HECLINET) ist eine Literaturdatenbank in deutscher und englischer Sprache auf dem Gebiet des Krankenhauswesens. In Kürze soll die gesamte MEDLINE-Datenbank durch das IR-System zugänglich gemacht werden und somit mehr für die Dermatologie relevante wissenschaftliche Informationen zur Verfügung stehen. Bezüglich der Integration von medizinischen Nachschlagewerken ist die Integration des Wörterbuches *Pschyrembel*[5] sowie eines dermatologischen Standardwerkes geplant. Abbildung 6.2 gibt einen Überblick über die verschiedenen Datenquellen, die über das IR-System per Freitextsuche zugänglich sind.

6.2.3 Benutzerinterface

Beim Benutzerinterface wurde größtmöglicher Wert auf einfache Bedienbarkeit der Suche gelegt. Das Eingabefeld, in dem ein Benutzer Freitext-Anfragen über alle im IR-System enthaltenen Informationen stellen kann, steht im Mittelpunkt der Seite. Zugeschnitten auf die ärztlichen Bedürfnisse werden einige zusätzliche Felder angeboten, wie beispielsweise die Patienten-ID, der Name eines Patienten oder sein Geschlecht. Die Ergebnisse einer Suchanfrage können nach Relevanz, nach Datum und

[2] `http://www.heydt.com/hydmedia.html`

[3] `http://www.dimdi.de/static/de/db/dbinfo/sm78.htm`, eingesehen im Februar 2007

[4] `http://www.dimdi.de/static/de/db/dbinfo/hn69.htm`, eingesehen im Februar 2007

[5] `http://www.pschyrembel.de`, eingesehen im Februar 2007

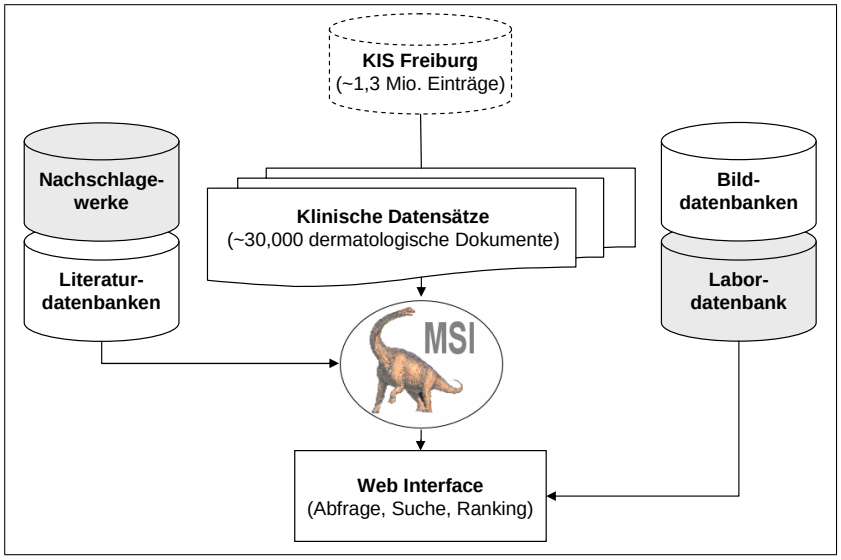

Abbildung 6.2: Übersicht über die verschiedenen Datenquellen, die über das Web-Interface zugänglich sind. Sämtliche textuellen Darstellungen werden durch das MorphoSaurus-System zunächst morpho-semantisch normalisiert. Strukturierte Daten und grafische Daten hingegen werden nicht vorverarbeitet. Grau unterlegt sind Datenquellen, die noch durch das IR-System zugänglich gemacht werden müssen.

nach Patientennamen sortiert werden. Bei den Ergebnissen wird jeweils der relevante Teil der entsprechenden Dokumente angezeigt. Suchwörter und ihre Variationen werden fett dargestellt. Zu jedem Dokument werden zusätzlich zu den interessanten Patientenstammdaten auch der Ersteller des Dokumentes und Erstellungsdatum angegeben. Gibt es zu einem gefundenen Dokument zugehörige Bildaufnahmen, so können diese per Mausklick direkt betrachtet werden. Abbildungen 6.3 und 6.4 zeigen Screenshots des IR-Systems. In Abbildung 6.3 sind in der Ergebnismenge die ersten drei Treffer zu sehen. Bei Mausklick auf das Auge-Icon am linken Rand des Screenshots auf Höhe des ersten Treffers öffnet sich ein Fenster (Abbildung 6.4) mit zugehörigen relevanten Bildern dieses Dokumentes.

6.3 Evaluation

Anhand einer Nutzerbefragung, an der alle verfügbaren Ärzte der Hautklinik Freiburg beteiligt waren, sollte der Nutzen des IR-Systems für die ärztliche Tätigkeit festgestellt werden. Hierbei sollte die Nützlichkeit für die klinische Arbeit, für Wissenschaft und Forschung sowie für die Lehre getrennt evaluiert werden. Den Anwendern wurde hierzu nach einer gewissen Eingewöhnungsphase ein Fenster im IR-System angezeigt, welches die in Listing 6.1 aufgeführten Fragen enthielt.

Abbildung 6.3: Benutzerinterface der MorphoSaurus-Suche in der Hautklinik Freiburg

Abbildung 6.4: Anzeige der zugehörigen Bilder zu den Resultaten aus der MorphoSaurus-Suche

Listing 6.1: Fragebogen zur Evaluation von MorphoSaurus in der Hautklinik

```
- Fragebogen -

1. klinische Arbeit
        a) Kann das System die Qualität der Behandlung verbessern?
        b) Kann das System helfen, Zeit zu sparen?
        Welche Einsatzszenarien sehen Sie für das System?
        c) Zu einem Fall vergleichbare Fälle finden?
        d) Training (z.B. Finden guter Arztbriefe)?
        e) Verbesserung der DRG-Kodierung?
        f) Ihre Vorschläge

2. wissenschaftliche Arbeit, Therapiestudien
        a) Kann das System die Qualität Ihrer Forschung verbessern?
        b) Kann das System helfen, Zeit zu sparen?
        c) Kann Sie das System unterstützen, Ihre Forschungs- und Studienprojekte
                besser oder präziser durchzuführen?
        d) Wenn ja, wie müsste es erweitert werden?
        Welche Einsatzszenarien sehen Sie für das System?
        e) Systematische Suche nach Krankheitsbildern
                (z.B. im Rahmen von Doktorarbeiten)?
        f) Therapievergleichsstudien?
        g) Identifizierung bestimmter Fallkonstellationen?
        h) Genotyp-/Phänotyp-Vergleich?
        i) Identifizierung seltener Arzneimittelnebenwirkungen?
        j) Ihre Vorschläge

3. Lehre
        a) Kann das System die Qualität Ihrer Lehrveranstaltungen verbessern?
        b) Kann das System helfen, Zeit zu sparen?
        Welche Einsatzszenarien sehen Sie für das System?
        c) Identifizierung interessanter Fälle für die Lehre?
        d) Finden charakteristischer Photos für die Lehre?
        e) Isolierung typischer Fälle (z.B. für Lehrbücher)?
        f) Ihre Vorschläge

4. Benutzeroberfläche
        a) Wie beurteilen Sie die Qualität der Treffer?
        b) Wie beurteilen Sie den Nutzen der Bildintegration?
        c) Wie empfinden Sie die Benutzeroberfläche?
        d) Welche zusätzlichen Funktionen wünschen Sie sich?
        e) Sonstige Angaben?
```

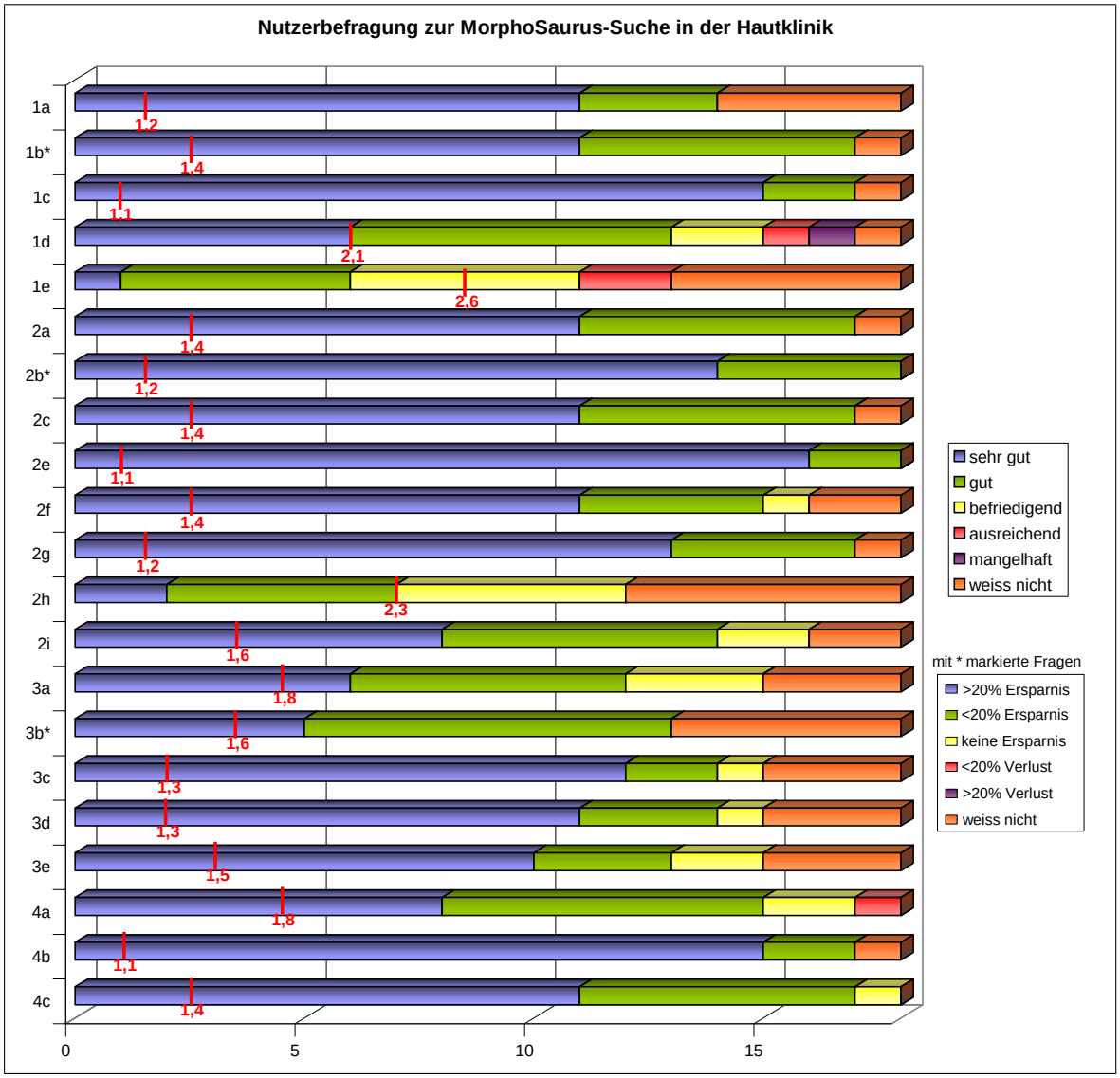

Abbildung 6.5: Übersicht über die Ergebnisse der Benutzerbefragung zur MorphoSaurus-Suche an der Hautklinik. Insgesamt nahmen 18 Ärztinnen und Ärzte an der Umfrage teil. Die meisten Fragen konnten mit Noten zwischen *sehr gut* und *mangelhaft* bewertet werden. Bei den mit Sternchen gekennzeichneten Fragen, bei denen nach der Zeitersparnis gefragt wurde, waren die Angaben *≥20% Zeitersparnis, <20% Zeitersparnis, keine Zeitersparnis, <20% Zeitverlust, ≥20% Zeitverlust* möglich. Rot angegeben sind die jeweiligen Durchschnittswerte pro Frage.

Das IR-System wurde von 18 Mitarbeitern (15 Ärzte, 3 Doktoranden) der Hautklinik Freiburg evaluiert. Die Akzeptanz dieses IR-Systems war überraschend hoch, wie der Ergebnisse in Abbildung 6.5 verdeutlichen. Erstmalig hatten die Ärzte Zugriff auf Tausende von Dokumenten und Bildern, die unabhängig von den einzelnen Patientenakten durchsuchbar waren. 17 von 18 Personen gaben an, dass das System eine erhebliche Erleichterung für ihre tägliche klinische Arbeit darstellt. Alle 18 Personen sehen die Zeitersparnis bei der wissenschaftlichen Arbeit als positiven Effekt des IR-Systems. Für die Lehre sehen 13 von 18 Personen einen positiven Nutzen des Systems (bei 5 Enthaltungen). Als besonderen Vorteil sehen die Befragten, dass sie nun im Rahmen von Forschungsprojekten systematisch ein bestimmtes Patientenkollektiv identifizieren können (Fragen 1c, 2e). Außerdem ist die Verknüpfung von Arztbriefsuche und Bildsuche ein entscheidender Pluspunkt der Anwendung (Frage 4b). Für drei Doktoranden wurde das System von Beginn an zu einem unersetzlichen Werkzeug für ihre Dissertationen. Weniger interessant erscheint die Anwendung, jedenfalls aus ärztlicher Sicht, für die DRG-Kodierung (1e) sowie für den Genotyp/Phänotyp-Vergleich (2h).

Bei den Fragen, für die eine Freitextantwort vorgesehen war (Frage 1f, 2j, 3f, 4d, 4e), zeigte sich, dass die Benutzer generell noch mehr Daten aus heterogenen Datenquellen wie Radiologieberichte oder Laborbefunde direkt per Mausklick abrufen möchten.

6.4 Diskussion der Ergebnisse und Ausblick

Unsere Erfahrungen zeigen, dass eine Verflechtung klinischer Informationssysteme mit Text-Mining-Methoden zu dem Zweck, sowohl strukturierte als auch unstrukturierte Daten durch ein leicht bedienbares Interface zur Verfügung zu stellen, ein Desideratum vieler Ärzte darstellt. Dies gilt sowohl für klinische Daten mit dem Ziel der vertikalen, also patientenübergreifenden, bereichsorientierten Sicht auf die Daten, aber auch für nicht patientenbezogene Daten wie wissenschaftliche Primärliteratur.

Trotz der hohen Akzeptanz unterliegt das System derzeit noch einigen Einschränkungen, die in den nächsten Monaten aufgehoben werden sollen:

- Die klinischen Daten sind bisher noch nicht in Echtzeit verfügbar, sondern werden aus den ursprünglichen Systemen an eine zentrale Stelle kopiert, auf welches das MorphoSaurus-System zugreift. Wünschenswerterweise sollten alle Daten schon ab der Entstehung in der Suche verfügbar sein.

- Das MorphoSaurus-Lexikon sollte spezifisch auf das dermatologische Voka-

bular trainiert werden, um die Qualität der Suche weiter zu steigern.

- Die Benutzerevaluation wurde bisher nur mit einem kleinen Kollektiv an Benutzern durchgeführt. Ausführliche Untersuchungen, die die Bedarfsanalyse der Ärzte und den Nutzen eines solchen Systems evaluieren, stehen noch aus.
- Eine Erweiterung des Systems auf andere Fachabteilungen innerhalb der Universitätsklinik Freiburg wird angestrebt.

Kapitel 7

Diskussion und Ausblick

In dieser Arbeit wurde das MorphoSaurus-System vorgestellt, welches Lösungsansätze für den Umgang mit zahlreichen linguistischen Variationen in der Medizin bietet. Diese Variationen erschweren die Entwicklung von sprachverarbeitenden Anwendungen in Klinik und Forschung, welche aufgrund des enorm steigenden Informationswachstums und -bedarfs in der Medizin eine immer wichtigere Rolle spielen. Innovative Sprachtechnologien wie das MorphoSaurus-System bilden somit eine notwendige Grundvoraussetzung im Umgang mit den Herausforderungen des modernen Gesundheitswesens. Die Vorstellung des MorphoSaurus-Systems beginnt bei seiner allgemeinen Architektur, geht über zu verschiedenen Anwendungen zur Verarbeitung medizinischer Texte und beschreibt schließlich die Integration von MorphoSaurus als ein intuitiver Suchdienst in das Informationssystem der Hautklinik Freiburg.

In der MorphoSaurus-Architektur werden die Struktur und das grundlegende Modell des MorphoSaurus-Systems näher beschrieben. Darin werden die lexikalischen Ressourcen definiert, die aus Subwort-Lexika und Thesaurus bestehen. Die Lexika enthalten mit den Subwörtern die semantisch elementaren Bausteine des MorphoSaurus-Systems. Über verschiedene Relationen sind diese Bausteine innerhalb des Thesaurus mehrsprachig miteinander verknüpft. Lexika und Thesaurus bilden die Grundlage des Zerlegungsalgorithmus, der dazu dient, Wörter in ihre semantischen Bestandteile zu zerlegen. Der Prozess der Zerlegung in semantisch atomare Einheiten wird als morpho-semantische Normalisierung bezeichnet. Das MorphoSaurus-Modell wurde in den vergangenen sechs Jahren für die biomedizinische Domäne implementiert.

Die morpho-semantische Normalisierung abstrahiert von linguistischen Variationen, die beispielsweise durch Flexion, Derivation und Komposition auftreten. Folgerichtig ist das MorphoSaurus-System in erster Linie für Anwendungen geeignet,

in denen linguistische Variationen keine maßgebliche Rolle für die Funktionalität einer Anwendung spielen. Dies ist im Information Retrieval in besonderem Maße gegeben. Ein Schwerpunkt dieser Arbeit liegt daher auf der Evaluation des MorphoSaurus-System in der Dokumentenrecherche. Hierbei konnte gezeigt werden, dass eine Kombination des MorphoSaurus-Systems mit klassischen regelbasierten Verfahren zu deutlichen Performanzgewinnen führt. Gleichzeitig wurde bei der Evaluation deutlich, dass die Retrievalperformanz entscheidend von der Qualität und der Abdeckung der lexikalischen Ressourcen abhängt. Nicht definierte oder zu unscharfe Äquivalenzklassen führen zu einer deutlichen Abnahme der Retrievalergebnisse. Es zeigte sich, dass die bestehenden lexikalischen Einträge im MorphoSaurus-System auf ihre semantische Schärfe hin zu überprüfen und dass insbesondere Lösungen für den Umgang mit Mehrwort-Einträgen zu erarbeiten sind. Als alleiniges Verfahren ist das MorphoSaurus-Verfahren derzeit nicht generell den regelbasierten Verfahren zur Stammformbildung wie dem Porter-Stemmer überlegen. Interessant sind die Ergebnisse, dass eine Kombination von Porter-Stemmer und MorphoSaurus-System im nicht-medizinischen Kontext der GIRT-Kollektion (Sozialwissenschaften) die größten Performanzgewinne gegenüber der Baseline erzielt (48,9% Zugewinn). Dies gibt Anlass zu der Vermutung, dass sich das MorphoSaurus-System auch in anderen Domänen außerhalb der Medizin sinnvoll in der Dokumentenrecherche einsetzen lässt.

Bei der automatischen Textkategorisierung von medizinischen Freitexten auf das englische MeSH-Vokabular zeigte sich, dass das verwendete statistische Verfahren dem regelbasierten Verfahren deutlich überlegen ist. Allerdings trägt auch das regelbasierte Verfahren zu einem Anstieg der Indexierungsergebnisse im kombinierten Verfahren bei. Die Gründe für das moderate Abschneiden der heuristischen Indexierung wurden ausführlich diskutiert, als ursächliche Faktoren wurden der Verlust linguistischer Informationen durch die morpho-semantische Normalisierung, fehlende Abdeckung der lexikalischen MorphoSaurus-Ressourcen, der problematische Gold-Standard sowie die fehlende Integration von an der NLM entwickelten Leitlinien und vorhandenem Expertenwissen genannt. IR-Experimente unter Hinzunahme automatisch ermittelter MeSH-Terme stehen derzeit noch aus, allerdings zeigt ein Großteil der Studien allenfalls einen moderaten Performanzgewinn in den Retrievalergebnissen.

Das MorphoSaurus-System wurde außerdem bei der automatischen Termübersetzung eingesetzt und zeigt für die deutsch-englische Übersetzung eine Performanz von 78,9% richtig oder in Beziehung stehender Übersetzungen. Die durchgeführte Fehleranalyse verdeutlichte, dass die Termübersetzung ganz entschei-

dend von den lexikalischen Ressourcen des MorphoSaurus-Systems abhängt. Da die Anzahl der Einträge in den lexikalischen Ressourcen des MorphoSaurus-Systems um ein Vielfaches geringer ausfällt als von herkömmlichen Vollformen-Lexika, bietet dieses Verfahren einen vielversprechenden Ansatz zur Übersetzung von Mehrwort-Ausdrücken und *Out-Of-Vocabulary*-Terms. Es wurde bereits erfolgreich bei der mehrsprachigen Dokumentenrecherche, bei der mit Hilfe dieses Verfahrens Benutzeranfragen in die Zielsprache übersetzt wurden [Daumke et al.2007b]. Außerdem zeigte sich, dass sich das Verfahren gut als Diagnosetool zur Validierung der lexikalischen Ressourcen von MorphoSaurus eignet. Es kann sprachübergreifend überprüft werden, welche Wörter auf einzelne oder mehrere Äquivalenzklassen abgebildet und somit fehlerhafte Abbildungen identifiziert werden.

Schließlich wurde das MorphoSaurus-System in Zusammenarbeit mit der Hautklinik Freiburg in das bestehende Informationssystem integriert. Es ermöglicht den intuitiven Zugriff auf einen Großteil der in der Hautklinik verfügbaren Daten über eine einzige Web-basierte Oberfläche. Das System wurde von Beginn an gut aufgenommen und wird bereits intensiv verwendet. In Zukunft sollen weitere Funktionalitäten wie der Abruf von Labordaten oder auch die Einbindung wissenschaftlicher Primär- und Sekundärliteratur erfolgen.

Zusammenfassend lässt sich sagen, dass die mit dem MorphoSaurus-System erzielten Ergebnisse durchaus ermutigend sind. Sie verdeutlichen gleichzeitig, dass das MorphoSaurus-System in vielfacher Hinsicht erweitert und qualitativ verfeinert werden kann. Hierzu zählen:

- Verbesserung der Qualität und der Abdeckung der bestehenden MorphoSaurus-Ressourcen.
- Aufnahme neuer Sprachen wie Holländisch oder Italienisch in die MorphoSaurus-Lexika.
- Bessere Unterstützung von Mehrwort-Einträgen. In diesem Zusammenhang sind Strategien zur Einbindung bestehender lexikalischer Ressourcen wie MeSH, UMLS oder ICD zu entwickeln.
- Optimierung der verwendeten Suchmaschinen wie Lucene auf das MorphoSaurus-System. Beispielsweise sind Techniken wie die Berücksichtigung von Nähebeziehung oder *Blind Relevance Feedback* in das bestehende System einzubinden.
- Einbindung weiterer Funktionalitäten zur Verarbeitung natürlicher Sprache. Insbesondere die Verwendungen von Wortarten-, Phrasen- und Namenserken-

nung kann in einigen Anwendungen wie der automatischen Textkategorisierung hilfreich sein.

- Ausführlicher *Proof-Of-Value* an der Hautklinik Freiburg: Studien belegen die Notwendigkeit intuitiver Suchfunktionalitäten für klinische und nicht-klinische Daten ("Google für Ärzte"). Erste Benutzertests zeigten die Begeisterung der Ärzte für das neue System. Dieses System gilt es jetzt mit weiteren Funktionalitäten auszustatten und ausführlich zu evaluieren.

Die notwendigen Schritte zur Verfeinerung des MorphoSaurus-Systems werden in der kommenden Zeit konsequent weiter verfolgt. Für rechnergestützte NLP-Anwendungen der Biomedizin leistet das MorphoSaurus-System jedoch schon jetzt einen wichtigen Beitrag.

Literaturverzeichnis

[Airio2006] Airio, Eija (2006). Word normalization and decompounding in mono- and bilingual IR. *Information Retrieval*, 9(3):249–271.

[AMIA1994] AMIA (1994). Standards for medical identifiers, codes, and messages needed to create an efficient computer-stored medical record. *Journal of the American Medical Informatics Association*, 1(1):1–7.

[Antman et al.1992] Antman, Eliot M., Joseph Lau, Bruce Kupelnick, Frederick Mosteller & Thomas C. Chalmers (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. *Journal of the American Medical Association*, 268(2):240–248.

[Apté et al.1994] Apté, Chidanand, Fred Damerau & Sholom M. Weiss (1994). Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems*, 12(3):233–251.

[Arampatzis et al.2000] Arampatzis, Avi, Theo P. van der Weide, Cornelis H. A. Koster & Patrick van Bommel (2000). Linguistically motivated information retrieval. In Allen Kent (Ed.), *Encyclopedia of Library and Information Science*, Vol. 69. Marcel Dekker, Inc., New York, Basel.

[Aronson et al.1999] Aronson, A. R., O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch & W. John Wilbur (1999). The indexing initiative. In *A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications.*

[Aronson2006] Aronson, Alan R. (2006). Metamap: Mapping text to the UMLS Metathesaurus.

[Aronson et al.2000] Aronson, Alan R., O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch & W. J. Wilbur (2000). The

NLM indexing initiative. In J. Marc Overhage (Ed.), *AMIA 2000 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care*, pp. 17–21. Los Angeles, CA, November 4-8, 2000. Philadelphia, PA: Hanley & Belfus.

[Aronson et al.2004] Aronson, Alan R., James G. Mork, Clifford W. Gay, Susanne M. Humphrey & Willie J. Rogers (2004). The NLM indexing initiative's medical text indexer. In Marius Fieschi, Enrico Coiera & Yu-Chan Jack Li (Eds.), *MEDINFO 2004 – Proceedings of the 11th World Congress on Medical Informatics. Vol. 1*, Studies in Health Technology and Informatics 107, pp. 268–272. San Francisco, CA, USA, September 7-11, 2004. Amsterdam: IOS Press.

[Aronson & Rindflesch1997] Aronson, Alan R. & Thomas C. Rindflesch (1997). Query expansion using the umls metathesaurus. In *AMIA'97 – Proc. of the 1997 AMIA Annual Fall Symposium (formerly SCAMC); Nashville, TN, 25–29 Oct 1997*, pp. 485–489.

[Aronson et al.1994] Aronson, Alan R., Thomas C. Rindflesch & Allan C. Browne (1994). Exploiting a large thesaurus for information retrieval. In *Proceedings of the RIAO 94 Conference: Computer-Assisted Information Searching on Internet*, pp. 197–216.

[Baeza-Yates & Ribeiro-Neto1999] Baeza-Yates, Ricardo & Berthier Ribeiro-Neto (Eds.) (1999). *Modern Information Retrieval*. Reading, MA: Addison-Wesley & Longman.

[Bauer1983] Bauer, Laurie (1983). *English Word-formation*. Cambridge, UK: Cambridge University Press.

[Brants2003] Brants, Thorsten (2003). Natural language processing in information retrieval. In *Proceedings of Computer Linguistics in the Netherlands*.

[Braschler & Ripplinger2004] Braschler, Martin & Bärbel Ripplinger (2004). How effective is stemming and decompounding for German text retrieval? *Information Retrieval*, 7(3-4):291–316.

[Brigl et al.1994] Brigl, Birgit, Markus Mieth, Reinhold Haux & Ewald Glück (1994). The LBI-method for automated indexing of diagnoses by using SNOMED. Part 1: Design and realization. *International Journal of Bio-Medical Computing*, 37(6):237–247.

[Brown et al.1990] Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

[Buckley et al.1994] Buckley, Chris, Gerard Salton, James Allan & Amit Singhal (1994). Automatic query expansion using SMART: TREC 3. In *Text REtrieval Conference*, pp. 69–80.

[Candler et al.2003] Candler, C. S., S. H. Uijtdehaage, & S. E. Dennis (2003). Introducing HEAL: The health education assets library. *Academic Medicine*, 78(3):249–253.

[CapGemini2005] CapGemini (2005). *Global research report Vision and Reality 2005.*

[Carpineto et al.2001] Carpineto, Claudio, Renato de Mori, Giovanni Romano & Brigitte Bigi (2001). An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27.

[Chen & Gey2004] Chen, Aitao & Fredric C. Gey (2004). Multilingual information retrieval using machine translation, relevance feedback and decompounding. *Information Retrieval*, 7(1-2):149–182.

[Cheng et al.2004] Cheng, Pu-Jen, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu & Lee-Feng Chien (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 146–153. New York, NY, USA: ACM Press.

[Chiao & Zweigenbaum2002] Chiao, Y.C. & P. Zweigenbaum (2002). Looking for french-english translations in comparable medical corpora. In Isaac S. Kohane (Ed.), *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical Informatics: One Discipline*, pp. 150–154. San Antonio, TX, November 9-13, 2002. Philadelphia, PA: Hanley & Belfus.

[Claveau & Zweigenbaum2005] Claveau, Vincent & Pierre Zweigenbaum (2005). Translating biomedical terms by inferring transducers. In *Artificial Intelligence in Medicine. Proceedings of the 10th Conference on Artificial Intelligence in Medicine in Europe – AIME 2005*, Vol. 3581, Lecture Notes in Artificial Intelligence, pp. 236–240. Aberdeen, Scotland, July 23 - 27, 2005. Berlin: Springer.

[Cooper & Miller1998] Cooper, Gregory F. & Randolph A. Miller (1998). An experiment comparing lexical and statistical methods for extracting mesh terms from clinical free text. *J Am Med Inform Assoc*, 5(1):62–75.

[Cooper & Byrd1997] Cooper, James W. & Roy J. Byrd (1997). Lexical navigation: visually prompted query expansion and refinement. In *DL '97: Proceedings of the second ACM international conference on Digital libraries*, pp. 237–246. New York, NY, USA: ACM Press.

[Côté et al.1993] Côté, Roger, David J. Rothwell, Ronald S. Beckett, James L. Palotay & Louise Brochu (1993). *The Systemised Nomenclature of MEDicine:* SNOMED *International*. Northfield, IL: College of American Pathologists.

[Crain1987] Crain, C (1987). Appendix A - protocol study of indexers at the national library of medicine. In Evans D. Scott D. Carbonell, J. & R. Thomason (Eds.), *Final Report on the Automated Classification Retrieval Project, Grant N01-LM-4-3529*. Bethesda, MD: National Library of Medicine.

[Dadam et al.2000] Dadam, Reichert & Kuhn (2000). Clinical workflows - the killer application for process-oriented information systems? In *Proceedings of the 4th International Conference on Business Information Systems*, pp. 36–59. Springer-Verlag.

[Daille et al.2000] Daille, Beatrice, Benoît Habert, Christian Jacquemin & Jean Royaute (2000). Empirical observation of term variations and principles for their description. *Terminology*, 3(2):197–257.

[Daumke et al.2007a] Daumke, Philipp, Kornél Markó, Jan Paetzold & Marcel Müller (2007a). Biomedical data mining in a hospital information system. In *MedNet 2007. To appear.*

[Daumke et al.2005a] Daumke, Philipp, Kornél Markó, Michael Poprat & Stefan Schulz (2005a). Multilingual biomedical dictionary. In *AMIA 2005 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, p. 933.

[Daumke et al.2007b] Daumke, Philipp, Kornél Markó, Michael Poprat, Stefan Schulz & Rüdiger Klar (2007b). Biomedical information retrieval across languages. *Informatics for Health and Social Care*, 32(2):131–147.

[Daumke et al.2006] Daumke, Philipp, Kornél Markó & Stefan Schulz (2006). Morphoogle - Eine multilinguale Suchmaschine für das WWW. In Klaus Haasis, Armin Heinzl & Dieter Klump (Eds.), *Aktuelle Trends in der Softwareforschung*, pp. 133–142. dpunkt.verlag.

[Daumke et al.2003] Daumke, Philipp, Kornél Markó, Stefan Schulz & Joachim Wermter (2003). Automatische MeSH-Indexierung auf der Basis morphosemantischer Normalisierung. In *GMDS 2003 – Tagungsband der 48. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, Vol. 34, pp. 225–228. Münster, Germany.

[Daumke et al.2005b] Daumke, Philipp, Stefan Schulz & Kornél Markó (2005b). A clir interface to a web search engine. In *AMIA 2005 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, p. 934.

[Daumke et al.2005c] Daumke, Philipp, Stefan Schulz & Kornél Markó (2005c). A clir interface to a web search engine. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Salvador, Brasil, August 15-19, 2005.

[Daumke et al.2005d] Daumke, Philipp, Stefan Schulz & Kornél Markó (2005d). Searching multilingual medical content in the web. *Technology and Health Care*, 13(5).

[Daumke & Markó2006] Daumke, P., Schulz St. & Kornél Markó (2006). Subword approach for acquiring and cross-linking multilingual specialized lexicons. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation Workshop: Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*.

[Dolin et al.2006] Dolin, Robert H., Liora Alschuler, Sandra Boyer, Calvin Beebe, Fred Behlen, Paul Biron & Amnon Shabo Shvo (2006). HL7 clinical document architecture, release 2. 13(1):30–39.

[Efthimiadis1996] Efthimiadis, Efthimis N. (1996). Query expansion. *Annual Review of Information Systems and Technology (ARIST)*, pp. 121–187.

[Fagan1989] Fagan, Joel L. (1989). The effectiveness of a nonsyntactic approach to automatic phrase indexing for document retrieval. *Journal of the American Society for Information Science*, 40:115–132.

[Fellbaum1998] Fellbaum, Christiane (Ed.) (1998). WordNet*: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

[Ferrucci & Lally2004a] Ferrucci, David & Adam Lally (2004a). Building an example application with the unstructured information management architecture. *IBM Systems Journal*, 43(3):455–475.

[Ferrucci & Lally2004b] Ferrucci, David & Adam Lally (2004b). Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

[Fluck1996] Fluck, Hans-Rüdiger (1996). *Fachsprachen: Einführung und Bibliographie* (5th ed.). Tübingen, Basel: Francke.

[Fung1998] Fung, Pascale (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber & Eduard H. Hovy (Eds.), *Machine Translation and the Information Soup. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas – AMTA 98*, Vol. 1529, Lecture Notes in Computer Science, pp. 1–17. Langhorne, PA, USA, October 28-31, 1998. Berlin: Springer.

[Fung & Yee1998] Fung, Pascale & Lo Yuen Yee (1998). An IR approach for translating new words from nonparallel, comparable texts. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 1, pp. 414–420. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

[Funk & Reid1983] Funk, Mark E. & Carolyn Anne Reid (1983). Indexing consistency in Medline. *Bulletin of the Medical Library Association*, 71(2):176–183.

[Furnas et al.1987] Furnas, George W., Thomas K. Landauer, Louis M. Gomez & Susan T. Dumais (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

[Gale et al.1992] Gale, William, Kenneth Ward Church & David Yarowsky (1992). Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, pp. 249–256. Morristown, NJ, USA: Association for Computational Linguistics.

[Gaudinat & Boyer2002] Gaudinat, Arnaud & Célia Boyer (2002). Automatic extraction of mesh terms from medline abstracts. In *NLPBA 2002, Workshop on Natural Language Processing in Biomedical Applications*.

[Gay et al.2005] Gay, Clifford W., Mehmet Kayaalp & Alan R. Aronson (2005). Semi-automatic indexing of full text biomedical articles. In *AMIA'05 – Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association*, pp. 271–275. Washington, D.C., November 22-26, 2003. Philadelphia, PA: Hanley & Belfus.

[Glatz-Krieger et al.2003] Glatz-Krieger, Katharina, Dieter Glatz, Margrith Gysel, Martina Dittler & Michael J. Mihatsch (2003). Webbasierte Lernwerkzeuge für die Pathologie - Web-based learning tools for pathology. *Pathologe*, 24:394–399.

[Gonzalo et al.1998] Gonzalo, Julio, Felisa Verdejo, Irina Chugur & Juan Cigarran (1998). Indexing with wordnet synsets can improve text retrieval. In *Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP*, pp. 38–44. Montreal, Canada.

[Grefenstette1994] Grefenstette, Gregory (1994). *Explorations in Automatic Thesaurus Discovery*. Norwell, MA, USA: Kluwer Academic Publishers.

[Harman1991] Harman, Donna (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15.

[Hayes & Weinstein1991] Hayes, Philip J. & Steven P. Weinstein (1991). CONSTRUE/TIS: A system for content-based indexing of a database of news stories. In *IAAI '90: Proceedings of the The Second Conference on Innovative Applications of Artificial Intelligence*, pp. 49–64. AAAI Press.

[Heckl2006] Heckl, Reiner W. (2006). *Sprachdummheiten in der Medizin* (3rd ed.). Darmstadt: Steinkopf Verlag.

[Helbig1998] Helbig, Hermann (1998). *Kurs 1699 - Automatische Sprachverarbeitung*. FernUniversität Hagen.

[Hersh et al.2005] Hersh, William, Jeffery Jensen, Henning Müller, Paul Gorman & Patrick Ruch (2005). A qualitative task analysis of biomedical image use and retrieval. In *ImageCLEF/MUSCLE workshop on image retrieval evaluation*, pp. 11–16.

[Hersh & Leone1995] Hersh, William & TJ Leone (1995). The saphire server: a new algorithm and implementation. In RM Gardner (Ed.), *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pp. 858–862.

[Hersh et al.2000] Hersh, William, Susan Price & Larry Donohoe (2000). Assessing thesaurus-based query expansion using the UMLS-metathesaurus. In J. Marc Overhage (Ed.), *AMIA 2000 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care*, pp. 344–348. Los Angeles, CA, November 4-8, 2000. Philadelphia, PA: Hanley & Belfus.

[Hersh2002] Hersh, William R. (2002). *Information Retrieval. A Health and Biomedical Perspective* (2nd ed.). New York: Springer.

[Hersh et al.1994a] Hersh, William R., Chris Buckley, T. J. Leone & David H. Hickam (1994a). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In Bruce Croft & C. J. van Rijsbergen (Eds.), *SIGIR'94 – Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192–201. Dublin, Ireland, 3-6 July 1994. London: Springer.

[Hersh & Dickham1994] Hersh, William R. & David H. Dickham (1994). The use of a multiapplication computer workstation in a clinical setting. *Bulletin of the Medical Library Association*, 82(4):382–389.

[Hersh & Donohoe1998] Hersh, William R. & Larry C. Donohoe (1998). SAPHIRE International: A tool for cross-language information retrieval. In C. G. Chute (Ed.), *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century*, pp. 673–677. Orlando, FL, November 7-11, 1998. Philadelphia, PA: Hanley & Belfus.

[Hersh & Hickam1995] Hersh, William R. & David H. Hickam, Hickam (1995). Information retrieval in medicine: The SAPHIRE experience. *Journal of the American Society for Information Science*, 46(10):743–747.

[Hersh et al.1994b] Hersh, William R., David H. Hickman, Brian Haynes & K. Ann McKibbon (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60.

[Hollink et al.2004] Hollink, Vera, Jaap Kamps, Christof Monz & Maarten De Rijke (2004). Monolingual document retrieval for European languages. *Information Retrieval*, 7(1-2):33–52.

[Hull1996] Hull, David A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84.

[Humphrey & Miller1987] Humphrey, Susanne M. & Nancy E. Miller (1987). Knowledge-based indexing of the medical literature: the indexing aid project. *Journal of the American Society fpr Information Science*, 38(3):184–196.

[Humphreys et al.1998] Humphreys, Betsy, Donald Lindberg, Harold Schoolman & G. Octo Barnett (1998). The Unified Medical Language System: An informatics research collaboration. *Journal of the American Medical Informatics Association*, 5(1):1–11.

[Ide & Véronis1998] Ide, Nancy & Jean Véronis (1998). Introduction to the Special Issue on Word Sense Disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

[Ide & Veronis1998] Ide, Nancy & Jean Veronis (1998). Word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

[Jacquemin et al.1997] Jacquemin, Christian, Judith L. Klavans & Evelyne Tzoukermann (1997). Expansion of multi-word terms for indexing and retrieval using morphology and syntax. In *Proceedings of the 35th annual meeting on Association for Computational Linguistics*, pp. 24–31. Morristown, NJ, USA: Association for Computational Linguistics.

[Jacquemin & Tzoukermann1999] Jacquemin, Christian & Evelyne Tzoukermann (1999). NLP for term variant extraction: Synergy between morphology, lexicon and syntax. In Tomek Strzalkowski (Ed.), *Natural language information retrieval*. Dordrecht: Kluwer Academic Publishers.

[Jenuwine & Floyd2004] Jenuwine, Elizabeth S. & Judith A. Floyd (2004). Comparison of Medical Subject Headings and text-word searches in MEDLINE to retrieve studies on sleep in healthy individuals. *Journal of the Medical Library Association*, 92(3):349–354.

[Joachims1999] Joachims, Thorsten (1999). Making large-scale support vector machine learning practical. pp. 169–184.

[Kantrowitz et al.2000] Kantrowitz, Mark, Behrang Mohit & Vibhu Mittal (2000). Stemming and its effects on *tfidf* ranking. In N. J. Belkin, P. Ingwersen & M.-K. Leong (Eds.), *SIGIR 2000 – Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 357–359. Athens, Greece, July 24-28, 2000. New York, NY: ACM.

[Kaplan1955] Kaplan, Abraham (1955). An experimental study of ambiguity and context. *Mechanical Translation*, 2(2):39–46.

[Kikui1998] Kikui, Genichiro (1998). Term-list translation using mono-lingual word co-occurrence vectors. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, pp. 670–374. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

[Kim et al.2001] Kim, Won, Alan R. Aronson & John Wilbur (2001). Automatic mesh term assignment and quality assessment. *AMIA'01 – Proceedings of the 2001 Annual Symposium of the American Medical Informatics Association*, pp. 319–323.

[Kluck2004] Kluck, Michael (2004). Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation. In J. Rittberger M. Bekavac, B. Herget (Ed.), *Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. Internationalen Symposiums für Informationswissenschaft (ISI 2004)*, pp. 247–268. Konstanz, Germany.

[Koch & Kaltenborn2005a] Koch, Oliver & Rossitza Kaltenborn (2005a). Das Arzt–Patienten-Verhältnis: Zwischen Individualisierung und Standardisierung. *Deutsches Ärzteblatt Online*.

[Koch & Kaltenborn2005b] Koch, Oliver & Rossitza Kaltenborn (2005b). Mehr Zeit für Patienten durch bessere Information. *Deutsches Ärzteblatt*, 28-29.

[Koskenniemi1984] Koskenniemi, Kimmo (1984). A general computational model for word formation recognition and production. In *COLING'84 – Proceedings of the 10th International Conference on Computational Linguistics & 22nd Annual Meeting of the Association for Computational Linguistics*, pp. 178–181. Stanford, California, U.S.A., 2-6 July 1984.

[Koster1999] Koster, C. (1999). Normalization and matching in the DORO system. In *Proceedings of the BCS-IRSG*.

[Kraaij & Pohlmann1996] Kraaij, Wessel & Renée Pohlmann (1996). Viewing stemming as recall enhancement. In H.-P. Frei, D. Harman, P. Schäuble & R. Wilkinson (Eds.), *SIGIR'96 – Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 40–48. Zurich, Switzerland, August 18-22, 1996. New York, NY: Association for Computing Machinery (ACM).

[Krovetz1993] Krovetz, Robert (1993). Viewing morphology as an inference process. In R. Korfhage, E. Rasmussen & P. Willett (Eds.), *SIGIR'93 – Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–203. Pittsburgh, PA, USA, June 27 - July 1, 1993. New York, NY: ACM.

[Kullback & Leibler1951] Kullback, Solomon & Richard A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

[Lancaster1991] Lancaster, Frederick Wilfrid (1991). *Indexing and abstracting in theory and practice*. University of Illinois: Champaign.

[Levenshtein1966] Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

[Levow et al.2005] Levow, Gina-Anne, Douglas W. Oard & Philip Resnik (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal*, 41(3):523–547.

[Lezius et al.1998] Lezius, Wolfgang, Reinhard Rapp & Manfred Wettler (1998). A freely available morphological analyzer, disambiguator and context sensitive lemmatizer for German. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 2, pp. 743–748. Montréal, Quebec, Canada, August 10-14, 1998. San Francisco, CA: Morgan Kaufmann.

[Lindberg et al.1993] Lindberg, Donald, Betsy Humphreys & Alexa T. McCray (1993). The Unified Medical Language Ssystem. *Methods of Information in Medicine*, 32(4):281–291.

[Lovins1968] Lovins, Julie B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1/2):22–31.

[Lüking1994] Lüking, Stephanie (1994). Bibliographie zur Fachsprache der Medizin. In *Wörterbücher der Medizin. Beiträge zur Fachlexikographie*. Stephan Dressler and Burkhard Schaeder.

[Markó2007] Markó, Kornél (2007). *MorphoSaurus - Foundations of Subword Indexing and its Multilingual Applications*. to appear. Dissertation, Language and Information Engineering Lab, Jena University, 2007.

[Markó et al.2003] Markó, Kornél, Philipp Daumke, Stefan Schulz & Udo Hahn (2003). Cross-language MeSH indexing using morpho-semantic normalization. In Mark A. Musen (Ed.), *AMIA'03 – Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications*, pp. 425–429. Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus.

[Markó et al.2004] Markó, Kornél, Udo Hahn, Stefan Schulz, Philipp Daumke & Percy Nohama (2004). Interlingual indexing across different languages. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pp. 82–99. Avignon, France, 26-28 April 2004. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

[Markó et al.2005a] Markó, Kornél, Stefan Schulz & Udo Hahn (2005a). Morphosaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545.

[Markó et al.2005b] Markó, Kornél, Stefan Schulz & Udo Hahn (2005b). Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI 2005 – Proceedings of the 20th National Conference on Artificial Intelligence & IAAI'05 – Proceedings of the 17th Innovative Applications of Artificial Intelligence Conference*, pp. 1075–1080. Pittsburgh, Pennsylvania, USA, July 9-13, 2004. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press.

[Markó et al.2006] Markó, Kornél, Stefan Schulz & Udo Hahn (2006). Cross-lingual alignment of biomedical acronyms and their expansions. In *MIE 2006 – Proceedings of the 20th International Congress of the European Federation of Medical Informatics*. Maastricht, Netherlands, August 27 - 30, 2006. Amsterdam: IOS Press.

[Markó et al.2005c] Markó, Kornél, Stefan Schulz, Alyona Medelyan & Udo Hahn (2005c). Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 528–535. Salvador, Brazil, August 15-19, 2005. New York, NY: ACM.

[McCallum & Nigam1998] McCallum, Andrew & Kamal Nigam (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*.

[McCray1998] McCray, Alexa T. (1998). The nature of lexical knowledge. *Methods of Information in Medicine*, 37(4/5):353–360.

[McCray et al.1988] McCray, Alexa T., Allen C. Browne & D. L. Moore (1988). The semantic structure of neo-classical compounds. In R. A. Greenes (Ed.), *SCAMC'88 – Proceedings of the 12th Annual Symposium on Computer Applications in Medical Care*, pp. 165–168. Washington, D.C., November 1988. New York, N.Y.: EEE Computer Society Press.

[McNamee & Mayfield2004] McNamee, Paul & James Mayfield (2004). Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97.

[Melamed2000] Melamed, I. Dan (2000). Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

[Mitra et al.1997] Mitra, Mandar, Christopher Buckley, Amit Singhal & Claire Cardie (1997). An analysis of statistical and syntactic phrases. In *Proceedings of RIAO-97, 5th International Conference "Recherche d'Information Assistee par Ordinateur"*, pp. 200–214. Montreal, CA.

[Monz & de Rijke2001] Monz, Christof & Maarten de Rijke (2001). Shallow morphological analysis in monolingual information retrieval for Dutch. In *2th Workshop of the Cross-Language Evaluation Forum, CLEF*. C. Peters, M. Braschler, J. Gonzalo, and M. Kluck.

[Monz & Dorr2005] Monz, Christoph & Bonnie J. Dorr (2005). Iterative translation disambiguation for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 520–527. Salvador, Brazil, August 15-19, 2005. New York, NY: ACM.

[Moulinier et al.2001] Moulinier, Isabelle, J. Andrew McCulloh & Elizabeth Lund (2001). West Group at CLEF 2000: Non-English monolingual retrieval. In *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000*, Vol. 2069, Lecture Notes in Computer Science. Heidelberg, Germany: Springer Verlag.

[Mueller et al.2003] Mueller, Marcel L., Rahul Butta & Hans-Ulrich Prokosch (2003). Electronic discharge letters using the clinical document architecture (cda). *Studies in Health Technology and Informatics*, 95:824–828.

[Müller et al.2006a] Müller, Henning, Thomas Deselaers, Thomas Lehmann, Paul Clough & William Hersh (2006a). Overview of the imageCLEFmed 2006 medical retrieval and annotation tasks. In *Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006*, LNCS, p. to appear. Alicante, Spain.

[Müller et al.2006b] Müller, Henning, Christelle Despont-Gros, William Hersh, Jeffery Jensen, Christian Lovis & Antoine Geissbuhler (2006b). Health care professionals image use and search behaviour. In *Proceedings of the Medical Informatics Europe Conference (MIE 2006)*, pp. 11–16.

[Müller et al.2004] Müller, Henning, Antoine Rosset, Jean-Paul Vallee, Francois Terrier & Antoine Geissbuhler (2004). A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, 28:295–305.

[Müller et al.2007] Müller, Marcel, Kornél Markó, Philipp Daumke, Jan Paetzold, Arnold Roesner & Rüdiger Klar (2007). Biomedical data mining in clinical routine: Expanding the impact of hospital information systems. In *MEDINFO 2007 – Proceedings of the 11th World Congress on Medical Informatics. To appear.*

[Nelson et al.2002] Nelson, Stuart J., Tammy Powell & Betsy L. Humphreys (2002). The Unified Medical Language System (UMLS) project. *Encyclopedia of Library and Information Science*, pp. 369–378.

[Névéol et al.2005a] Névéol, Aurélie, Vincent Mary, Arnaud Gaudinat, Célia Boyer, Alexandrina Rogozan & Stéfan J. Darmoni (2005a). A benchmark evaluation of the French MeSH indexers. In *AIME'05 — Proceedings of the 10th Conference on Artificial Intelligence in Medicine*, pp. 251–255.

[Névéol et al.2005b] Névéol, Aurélie, James G. Mork, Alan R. Aronson & Stéfan J. Darmoni (2005b). Evaluation of French and English MeSH indexing systems with a parallel corpus. In *AMIA'05 – Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association*, pp. 565–569. Washington, D.C., November 22-26, 2005. Philadelphia, PA: Hanley & Belfus.

[Névéol et al.2006a] Névéol, Aurélie, Alexandrina Rogozan & Stéfan Darmoni (2006a). Automatic indexing of online health resources for a French quality controlled gateway. *Information Processing and Management*, 42(3):695–709.

[Névéol et al.2006b] Névéol, Aurélie, Kelly Zeng & Olivier Bodenreider (2006b). Besides precision and recall: Exploring alternative approaches to evaluating an automatic indexing tool for medline. *AMIA'06 – Proceedings of the 2006 Annual Symposium of the American Medical Informatics Association*, pp. 589–593.

[Niedermair et al.1984] Niedermair, G. T., G. Thurmair & I. Büttel (1984). MARS: A retrieval tool on the basis of morphological analysis. In C. J. van Rijsbergen (Ed.), *Proceedings of the 3rd Joint BCS and ACM Symposium on Research and Development in Information Retrieval*, pp. 369–381. Cambridge, England, 2-6 July, 1984. Cambridge, U.K.: Cambridge University Press.

[Nunzio et al.2004] Nunzio, Girgio M. Di, Nicola Ferro, Massimo Melucci & Nicola Orio (2004). Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In Carol Peters, Julio Gonzalo, Martin Braschler & Michael Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003*, pp. 220–235. Berlin, Springer Verlag.

[Oard1997] Oard, Douglas W. (1997). Alternative approaches for cross-language text retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval.* American Association for Artificial Intelligence.

[Oard2002] Oard, Douglas W. (2002). When you come to a fork in the road, take it: Multiple futures for CLIR research. In *SIGIR 2002 Workshop on the Future of Cross-Language Information Retrieval Research.* August 2002, Tampere, Finland.

[Oard & Diekema1998] Oard, Douglas W. & Anne R. Diekema (1998). Cross-language information retrieval. In Martha E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST), Vol. 33: 1998*, pp. 223–256. Medford, NJ: Information Today.

[Pirkola2001] Pirkola, Ari (2001). Morphological typology of languages for IR. *Journal of Documentation*, 57(3):330–348.

[Popovič & Willett1992] Popovič, Mirko & Peter Willett (1992). The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390.

[Porter1980] Porter, Martin F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

[Pouliquen et al.2002] Pouliquen, Bruno, Denis Delamarre & Pierre Le Beux (2002). Indexation de textes médicaux par extraction de concepts et ses utilisations. In *Journées internationales dAnalyse statistique des Données Textuelles*, 2, pp. 617–628.

[Price1963] Price, Derek (1963). *Little Science, Big Science*. New York: Columbia University Press.

[Rapp1999] Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 519–526. College Park, MD, USA, 20-26 June 1999. San Francisco, CA: Morgan Kaufmann.

[Resnik & Melamed1997] Resnik, P. & I. Melamed (1997). Semi-automatic acquisition of domain-specific translation lexicons. In *ANLP 1997 – Proceedings of the 5th Conference on Applied Natural Language Processing*, pp. 340–347. Washington, D.C., March 31 - April 3, 1997. San Francisco, CA: Morgan Kaufmann.

[Resnik1999] Resnik, Philip (1999). Mining the web for bilingual text. In *Proceedings of the International Conference of the Association of Computational Linguistics.*, pp. 527–534.

[Robertson1990] Robertson, Stephen E. (1990). On term selection for query expansion. *Journal of Documentation*, 46(4):359–364.

[Robertson & Soboroff2001] Robertson, Stephen E. & Ian Soboroff (2001). The TREC 2001 filtering track report. In *Text REtrieval Conference*.

[Robertson et al.2000] Robertson, Stephen E., S. Walker & M. Beaulieu (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*, 36(1):95–108.

[Rocchio1971] Rocchio, Joseph J. (1971). Relevance feedback in information retrieval. *The SMART retrieval system. Experiments in automatic document processing*, 27(3):313–323.

[Roelcke2005] Roelcke, Thorsten (2005). *Fachsprachen*. Berlin: Erich Schmidt Verlag.

[Ruch et al.2003] Ruch, Patrick, Robert H. Baud & Antoine Geissbühler (2003). Learning-free text categorization. In *AIME*, pp. 199–208.

[Sanderson1994] Sanderson, Mark (1994). Word sense disambiguation and information retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 142–151. New York, NY, USA: Springer-Verlag New York, Inc.

[Sanderson2000] Sanderson, Mark (2000). Retrieving with good sense. *Information Retrieval*, 2(1):49–69.

[Sapir1921] Sapir, Edward (1921). *Language: An Introduction to the Study of Speech*. Harcourt, Brace and Company.

[Savoy2003] Savoy, Jacques (2003). Report on CLEF 2002 experiments: Combining multiple sources of evidence. In *3th Workshop of the Cross-Language Evaluation Forum, CLEF*. C. Peters, M. Braschler, J. Gonzalo, and M. Kluck.

[Savoy2006] Savoy, Jacques (2006). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *SAC '06: Proceedings of the 2006 ACM symposium on Applied computing*, pp. 1031–1035. New York, NY, USA: ACM Press.

[MeSH2005] MeSH (2005). *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

[MeSH2006] MeSH (2006). *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

[Umls2005a] Umls (2005a). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

[Umls2005b] Umls (2005b). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

[Schapire & Singer2000] Schapire, Robert E. & Yoram Singer (2000). Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39(2-3):135–168.

[Schulz & Hahn2000] Schulz, Stefan & Udo Hahn (2000). Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics*, 59(3):87–99.

[Schulz et al.2005] Schulz, Stefan, Susanne Hanser, Udo Hahn & Rüdiger Klar (2005). Semantische klarstellung der repräsentation von prozeduren in snomed ct. In R. Klar, W Köpcke, K Kuhn, H. Lax, S. Weiland & A. Zaiss (Eds.), *GMDS 2005 – Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*. Freiburg, 11-15 September, 2005.

[Schulz et al.2002] Schulz, Stefan, Martin Honeck & Udo Hahn (2002). Biomedical text retrieval in languages with a complex morphology. In Stephen Johnson (Ed.), *Proceedings of the ACL/NAACL 2002 Workshop on 'Natural Language Processing in the Biomedical Domain'*, pp. 61–68. University of Pennsylvania, Philadelphia, PA, USA, July 11, 2002. New Brunswick, NJ: Association for Computational Linguistics (ACL).

[Schulz et al.2004] Schulz, Stefan, Kornél Markó, Eduardo Sbrissia, Percy Nohama & Udo Hahn (2004). Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, Vol. 2, pp. 813–819. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.

[Schütze & Pedersen1995] Schütze, Hinrich & Jan Pedersen (1995). Information retrieval based on word senses. In *Symposium on Document Analysis and Information Retrieval (SDAIR)*, pp. 161–175.

[Sebastiani2002] Sebastiani, Fabrizio (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

[Sheridan & Ballerini1996] Sheridan, Paraic & Jean Paul Ballerini (1996). Experiments in multilingual information retrieval using the SPIDER system. In H.-P. Frei, D. Harman, P. Schäuble & R. Wilkinson (Eds.), *SIGIR'96 – Proceedings of*

the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 58–65. Zurich, Switzerland, August 18-22, 1996. New York, NY: Association for Computing Machinery (ACM).

[Smeaton1999] Smeaton, Alan F. (1999). Using NLP or NLP resources for information retrieval tasks. In Tomek Strzalkowski (Ed.), *Natural language information retrieval*. Dordrecht: Kluwer Academic Publishers.

[Smeaton et al.1995] Smeaton, Alan F, Fergus Kelledy & Ruairi O'Donnell (1995). Thresholding posting lists, query expansions with WordNet and POS tagging of Spanish. In *Text REtrieval Conference*.

[Snomed CT2006] Snomed CT (2006). Snomed *Clinical Terms*. Northfield, IL: College of American Pathologists.

[Srinivasan1996a] Srinivasan, Padmini (1996a). Optimal document-indexing vocabulary for medline. *Information Processing and Management*, 32(5):503–514.

[Srinivasan1996b] Srinivasan, Padmini (1996b). Query expansion and medline. *Information Processing and Management*, 32(4):431–443.

[Stempfhuber & Baerisch2006] Stempfhuber, Maximilian & Stefan Baerisch (2006). Domain-specific track CLEF 2006: Overview of results and approaches, remarks on the assessment analysis. In *7th Workshop of the Cross-Language Evaluation Forum, CLEF 2006 (erscheint)*.

[Strzalkowski et al.1999] Strzalkowski, Tomek, Fang Lin, Jin Wang & Jose Perez-Carballo (1999). Evaluating natural language processing techniques for information retrieval: a TREC perspective. In Tomek Strzalkowski (Ed.), *Natural language information retrieval*. Dordrecht: Kluwer Academic Publishers.

[Tomlinson2001] Tomlinson, Stephen (2001). Stemming evaluated in 6 languages by Hummingbird SearchServer$^{\text{tm}}$ at CLEF 2001. In Carol Peters, Martin Braschler, Julio Gonzalo & Michael Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001*, Vol. 2406, pp. 278–287.

[Tomlinson2004] Tomlinson, Stephen (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer at CLEF 2003. In Carol Peters, Julio Gonzalo, Martin Braschler & Michael Kluck (Eds.),

Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003, pp. 286–300. Berlin, Springer Verlag.

[Tordai & de Rijke2005] Tordai, Anna & Maarten de Rijke (2005). Four stemmers and a funeral: Stemming in Hungarian at CLEF 2005. In Carol Peters (Ed.), *Working Notes for the 2005 CLEF Workshop*. Vienna, Austria, 21-23 September.

[UIMA2006] UIMA (2006). *Unstructured Information Management Architecture (UIMA). SDK User's Guide and Reference. Version 2.* Internation Business Machines Corporation (IBM).

[Ulrich2000] Ulrich, Volker (2000). Medizinisch–technischer Fortschritt, demographische Alterung und Wachstum der Gesundheitsausgaben: Was sind die treibenden Faktoren? *Gesundheitsökonomie und Qualitätsmanagement*, (5).

[Voorhees1994] Voorhees, Ellen M. (1994). Query expansion using lexical-semantic relations. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 61–69. New York, NY, USA: Springer-Verlag New York, Inc.

[Vossen1998] Vossen, Piek (Ed.) (1998). EuroWordNet*: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

[Wagner et al.1995] Wagner, Judith C., Danny Solomon, Pierre-André Michel, Robert H. Baud Christian Juge, Alan L. Rector AL & Jean-Raoul Scherrer (1995). Multilingual natural language generation as part of a medical terminology server. In R. A. Greenes, H. E. Peterson & D. J. Protti (Eds.), *MEDINFO'95 – Proceedings of the 8th Conference on Medical Informatics*, IFIP World Conference Series on Medical Informatics, pp. 100–104. Vancouver, Canada, 1995. Amsterdam: North-Holland.

[Wallis et al.1995] Wallis, Jerold W., Michelle M. Miller, Tom R. Miller & Thomas H. Vreeland (1995). An internet-based nuclear medicine teaching file. *Journal of Nuclear Medicine*, 36(8):1520–1527.

[Wessel Kraaij1998] Wessel Kraaij, Renee Pohlmann (1998). Comparing the effect of syntactic vs. statistical phrase indexing strategies for Dutch. In C. Nikolaou & C. Stephanidis (Eds.), *Research and Advanced Technology for Digital Libraries.*

Proceedings of the 2nd European Conference – ECDL'98, Lecture Notes in Computer Science 1513, pp. 605–616. Heraklion, Crete, Greece. Berlin, Heidelberg, New York: Springer.

[Winter et al.2002] Winter, Alfred, Elske Ammenwerth, Birgit Brigl & Reinhold Haux (2002). Krankenhausinformationssysteme. In *Handbuch der Medizinischen Informatik*. München/Wien: Carl Hanser.

[Won Kim2001] Won Kim, W. John Wilbur (2001). Corpus-based statistical screening for content-bearing terms. *Journal of the American Society for Information Science and Technology*, 52(3):247–259.

[Xu et al.2004] Xu, Hua, Kristin Anderson, Victor R. Grann & Carol Friedman (2004). Facilitating cancer research using natural language processing of pathology reports. In Marius Fieschi, Enrico Coiera & Yu-Chan Jack Li (Eds.), *MEDINFO 2004 – Proceedings of the 11th World Congress on Medical Informatics. Vol. 1*, Studies in Health Technology and Informatics 107, pp. 565–569. San Francisco, CA, USA, September 7-11, 2004. Amsterdam: IOS Press.

[Yang1999] Yang, Yiming (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1/2):69–90.

[Yiming Yang1994] Yiming Yang, Christopher G. Chute (1994). Words or concepts: the features of indexing units and their optimal use in information retrieval. In *Seventeenth Annual Symposium on Computer Applications in Medical Care*, pp. 685–689. Washington, D. C.: McGraw-Hill.

[Zhang & Vines2004] Zhang, Ying & Phil Vines (2004). Using the web for automated translation extraction in cross-language information retrieval. In *SIGIR 2004 – Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169. Sheffield, United Kingdom. New York, NY: ACM.

[Zweigenbaum et al.2001] Zweigenbaum, Pierre, Stéfan J. Darmoni & Natalia Grabar (2001). The contribution of morphological knowledge to French MeSH mapping for information retrieval. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past*, pp. 796–800. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Danksagung

An dieser Stelle möchte ich mich bei allen bedanken, die zum Gelingen dieser Arbeit beigetragen haben. Für die exzellente Betreuung und das mir entgegengebrachte Vertrauen möchte ich mich bei meinem Erstgutachter PD. Dr. Stefan Schulz bedanken. Herrn PD Dr. Elmar Kotter danke ich für die Bereitschaft, diese Arbeit als Zweitgutachter anzunehmen. Mein Dank gilt außerdem Herrn Prof. Klar für die großartige Unterstützung unserer Arbeitsgruppe.

Mein besonderer Dank gilt Kornél Markó, der mich während meiner gesamten wissenschaftlichen Arbeit intensiv unterstützt hat. Marcel Müller danke ich für die vorzügliche Zusammenarbeit bei der Integration von MorphoSaurus in die Hautklinik. Außerdem bedanke ich mich bei Jan Paetzold, der für den einwandfreien Ablauf unserer täglichen Arbeit sorgt und mir in allen Fragen zur Programmierung beiseite steht. Auch Michael Poprat hat mich in einigen Projekten unterstützt, wofür ich ihm herzlich danke. Edson Pacheco danke ich für seine wertvollen Tipps zur Java-Programmierung und Frau Richards für das Gegenlesen dieser Arbeit.

Auch bei allen "Morphosauriern", die in den letzten Jahren zum Erfolg des Projektes beigetragen haben, möchte ich mich sehr herzlich bedanken. Diese sind: Roosewelt Leite de Andrade, Andreas Baum, Ines Beiser, Jeferson Bitencourt, Emerson Borsato, Rafael Bruns, Pindaro Cancian, Nicolau Carboni, Dhayana Dallmeier, Bruno Duque, Adriano Duma, Ana Bravo Ferer, Claudia Fink, Guilherme del Fiol, Pius Franz, Florian Grund, Maria Claudia Hahn, Prof. Dr. Udo Hahn, Dr. Susanne Hanser, Anna-Karin Hermansson, Martin Honeck, Janine Kaliniak, Dörte Jensen, Gabriel Khalil, Graziele Klein, Pricilla Koppe, Martin Krüger, Thais Machado, Lucio Matias, Josiane Melchioretto, Paulo Mendes, Prof. Dr. Claudia M. Moro, Guilherme Neto, Prof. Dr. Percy Nohama, Michel Oleynik, Oliver Osburg, Luciana Ribeiro, Dr. Martin Romacker, Viviane Seki Sassaki, Ricardo César Ribeiro dos Santos, Eduardo Rocha Sbrissia, Eva Schulte, Michael Schultheiss, Martin Schwarz, Michelli Paula da Silva, Hood Wilson Gusso Silva, Anders Thurin, Anderson Trindade Venturini, Joachim Wermter, Oliver Würstlin und Albrecht Zaiss.

Ein besonderer Dank geht an meine Eltern, die mich auf meinem bisherigen Weg stets unterstützt haben und mir helfen, alle meine beruflichen Ziele zu verwirklichen. Insbesondere danke ich Ihnen für den Respekt zu meiner Entscheidung, nicht ein "normaler Arzt" zu werden.

Diese Arbeit wäre nicht entstanden ohne die unermessliche Liebe und Geduld meiner Frau Sarah, die mich immer unterstützt hat. Ihr kann ich nie genug dankbar sein.